DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1 Article 16

5-1-2003

Using Multinomial Logistic Models To Predict Adolescent Behavioral Risk

Chao-Ying Joanne Peng
Indiana University, peng@indiana.edu

Rebecca Naegle Nichols
Indiana University, rnaegle@yahoo.com

Recommended Citation

Peng, Chao-Ying Joanne and Nichols, Rebecca Naegle (2003) "Using Multinomial Logistic Models To Predict Adolescent Behavioral Risk," *Journal of Modern Applied Statistical Methods*: Vol. 2 : Iss. 1 , Article 16.
DOI: 10.22237/jmasm/1051748160
Available at: http://digitalcommons.wayne.edu/jmasm/vol2/iss1/16

Using Multinomial Logistic Models To Predict Adolescent Behavioral Risk

Cover Page Footnote
We wish to thank Gary M. Ingersoll for the use of the data.

Journal of Modern Applied Statistical Methods
May, 2003, Vol 2, No. 1, 177-188

Copyright © 2003 JMASM, Inc.
1538 – 9472/03/$30.00

Using Multinomial Logistic Models To Predict Adolescent Behavioral Risk

Chao-Ying Joanne Peng
School of Education
Indiana University-Bloomington

Rebecca Naegle Nichols
School of Health, Physical Ed. and Recreation
Indiana University-Bloomington

Multinomial logistic regression was applied to data comprising 432 adolescents' self reports of engagement in risky behaviors. Results showed that gender, intention to drop from the school, family structure, self-esteem, and emotional risk were effective predictors collectively. Three methodological issues were highlighted: (1) the use of odds ratio, (2) the absence of an extension of the Hosmer and Lemeshow test for multinomial logistic models, and (3) the missing data problem. Psychologists and educators can utilize findings to plan prevention programs, as well as to apply the versatile and effective logistic technique in psychological, educational, and health research concerning adolescents.

Key words: Adolescent behavior, self-esteem, behavioral risk, emotional risk, family structure, multinominal logistic model, logistic modeling

Introduction

Adolescence is a very influential time in the life of a young person. It is a time of change and possible insecurity, accompanied by low self-esteem and emphasis on peer approval (Bergman & Scott, 2001; Brack, Orr, & Ingersoll, 1988; McGee & Williams, 2000). This may be the reason that many risky health habits are developed during adolescence. One example is smoking. A study conducted by Everett and Husten (1999) revealed that 81% of college aged students who reported ever being daily smokers began smoking before the age of 18. Furthermore, they found that among those who ever smoked a whole cigarette, 43.0% did so for the first time at the age of 14 or younger; 23.7% at age 15 or 16. Other researchers have come to similar conclusions regarding the adoption of risky health habits during adolescents (Bergman & Scott, 2001; McGee & Williams, 2000; Orr, Wilbrandt, Brack, Rauch, & Ingersoll, 1989).

Because many health-endangering behaviors are engaged in for the first time during adolescence, one goal of health education is to reduce the initiation of health-endangering behaviors. These behaviors include, but are not limited to, unsafe sexual activity (Orr, et al., 1989) and the use of alcohol, tobacco, and marijuana (McGee & Williams, 2000). It is essential that health educators identify those youth at greatest risk so that effective programs may be targeted specifically toward minimizing or eliminating these behaviors. In this paper, we demonstrate the utility of multinomial logistic regression model in identifying adolescents at greatest health risk from their personal as well as family characteristics. Psychologists and educators can utilize findings to plan prevention programs, as well as to apply the versatile logistic regression technique in psychological, educational, and health research concerning adolescents.

Logistic regression is a promising statistical technique that can be used to predict the likelihood of a categorical outcome variable. It has found widespread use in the epidemiological literature, where often the dependent variable is presence or absence of a disease state. This technique has also proven useful in broader areas — social sciences (e.g., Chuang, 1997; Janik and Kravitz, 1994; Tolman and Weisz, 1995) and education, especially higher education (Austin,

Send correspondence to Joanne Peng, Department of Counseling and Educational Psychology, 4050 201 N. Rose Ave., Bloomington, IN 47405-1006 or email peng@indiana.edu. E-mail Rebecca Naegle at rnaegle@yahoo.com. We wish to thank Gary M. Ingersoll for the use of the data.

Yaffee, & Hinkle, 1992, Cabrera, 1994; Peng, So, Stage, & St. John, 2002) — than the typical epidemiological situation. To profile adolescents who are at greatest risk of participation in risky health behaviors, multinomial logistic regression was applied to data comprising 432 adolescents' self reports of engagement in risky behaviors. Results are interpreted in terms of substantive and methodological implications. The remainder of this paper is divided into four sections: (1) Methodology, (2) The Multinomial Logistic Regression Model, (3) Interpreting and Assessing Multinomial Logistic Regression Results, and (4) Conclusion.

Methodology

Self-reported health behavior data were collected from 517 adolescents enrolled in two junior high schools (grades 7 through 9) in the fall of 1988. Parents were notified by mail that the survey was to be conducted. Both the parents and the students were assured of their rights to optional participation and confidentiality of students' responses. Written parental consent was waived with the approval of the school administration and the university Institutional Review Board (Ingersoll, Grizzle, Beiter, & Orr, 1993). Among the 517 students, 85 did not complete all questions. Thus, the final sample size was 432 (83.4% were Whites and the remaining Blacks or others) with a mean age of 13.9 years and nearly even composition of girls (*n*=208) and boys (*n*=224). The problem with missing data is addressed later in a section titled Missing Data.

Health Behavior Questionnaire (HBQ; Ingersoll & Orr, 1989; Resnick, Harris, & Blum, 1993) and Rosenberg's self esteem inventory (Rosenberg, 1965) were administered on the same day to all students in all math classes (a mandatory subject). The HBQ asked adolescents to indicate whether they engaged in specific risky health behaviors (Behavioral Risk Scale) or had experienced selected emotions (Emotional Risk Scale). Examples of behavioral risk items were "I use alcohol (beer, wine, booze)," "I use pot," and "I have had sexual intercourse/gone all the way." These items measured frequency of adolescents' alcohol and drug use, sexual activity, and delinquent behavior. They were responded to on a 4-point ordinal scale (1=never, and 4=about once a week). Emotional risk items measured adolescents' quality of relationship with others, and management of emotions (e.g., "I have attempted suicide," "I have felt depressed," etc.). Cronbach's alpha reliability (Nunnally, 1977) was 0.84 for the Behavioral Risk Scale and 0.81 for the Emotional Risk Scale.

Adolescents' self esteem was assessed using Rosenberg's self esteem inventory (Rosenberg, 1965). Self-esteem scores ranged from 9.79 to 73.87 with a mean of 49.97 and standard deviation of 10.09. Furthermore, among the 432 adolescents, 12.27% (or 53) indicated an intention to drop out of school; 44.68% (or 193) were from intact families, 22.69% (or 98) were from families with one step-parent, and 32.63% (or 141) were from families headed by a single parent.

For the present data, we were interested in identifying adolescents at the greatest behavioral risk from their gender, intention to drop out from school, family characteristics, emotional risks, and self-esteem scores. In addition to identifying youth at the greatest behavioral risk, we were also interested in differentiating adolescents at medium level of risk from those at low risk so that psychologists and educators could utilize findings to design appropriate prevention programs to help adolescents with different needs. Given the objective of this study, the research hypothesis posed to the data was stated as follows: "the likelihood that an adolescent is at high, medium, or low behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self esteem." The dependent variable was students' risk level on the Behavioral Risk Scale of the HBQ; it is hereafter referred to as the RISK variable. The explanatory variables included gender, intention to drop out of school, type of family structure, emotional risk, and self-esteem scores.

Scores on the Behavioral Risk Scale of the HBQ ranged from 40.44 to 66.81 with a mean of 47.69 and a standard deviation of 10.89. Adolescents at highest behavioral risk (*n*=29) were identified to be those scored at least one standard deviation above the mean, i.e., 60 or higher. Those scored between 45 and 59 were identified to be at medium behavioral risk (*n*=170), and those scored between 44 and 40 were at low behavioral risk (*n*=233). The cutoff used to separate those at medium risk from those at low risk was the median of the distribution (between 44 and 45), given the positive skewness of the scores on the Behavioral Risk Scale and the 4-point scale used for each item. Those classified as at low behavior risk were adolescents who answered, on the average, between "never", coded as 1, and "once in a while", coded as 2.

The relationship between the RISK dependent variable and each of the three categorical explanatory variables is shown in Tables 1A through 1C. According to Table 1A, boys were classified into high or medium behavioral risk groups more frequently than girls while the trend was reversed for the low risk group. Table 1B revealed that adolescents intending to drop out of school were more likely to exhibit high or medium behavioral risk than those without such an intention. As for the relationship between family structures and behavioral risk, a majority of adolescents from either intact or step-parent families exhibited a low level of behavioral risk whereas a majority of those from single-parent families showed a medium level of behavioral risk (Table 1C).

Table 1A. Distribution of Gender and Three Levels of Behavioral Risk.

Behavioral Risk	Gender		Total
Levels	Girls=0	Boys=1	
High Risk	5	24	29
Medium Risk	66	104	170
Low Risk	137	96	233
Total	208	224	432

Table 1B. Distribution of Dropout and Three Levels of Behavioral Risk.

Behavioral Risk	Dropout		Total
Levels	No=0	Yes=1	
High Risk	15	14	29
Medium Risk	137	33	170
Low Risk	227	6	233
Total	379	53	432

Table 1C. Distribution of Family Structure and Three Levels of Behavioral Risk.

Behavioral Risk	Family Structure			Total
Levels	Intact=1	Step=2	Single=3	
High Risk	8	7	14	29
Medium Risk	62	38	70	170
Low Risk	123	53	57	233
Total	193	98	141	432

The Multinomial Logistic Regression Model

Logistic regression is well suited for describing and testing hypotheses about relationships between a categorical dependent variable and one or more categorical or continuous explanatory variables. Specifically, multinomial logistic regression was chosen to answer the research question for two reasons. First, multinomial logistic regression provides an effective and reliable way to obtain the estimated probability of belonging to a specific population (e.g., high risk adolescents) and the estimate of odds ratio of adolescents' characteristic on their behavioral risk (Peng, Lee, & Ingersoll, 2002; Peng, Manz, & Keck, 2001; Scott, Mason, & Chapman, 1999).

Second, multinomial logistic regression is a procedure by which estimates of the net effects of a set of explanatory variables on the dependent variable can be obtained (Morgan & Teachman, 1988). Even though logistic regression has been used in health research, the use of multinomial logistic regression is rare. In this section, we will first describe the general logic behind the multinomial logistic regression model. This is followed by the specification of a multinomial logistic model for the present data in order to answer the research question.

The simplest form of the multinomial logistic regression model involves one categorical dependent variable *Y* (e.g., three levels of behavioral risk) and one explanatory variable, *X* (e.g., emotional risk score). Let p_1= the probability of high behavioral risk (Y=3), p_2= the probability of medium behavioral risk (Y=2), and p_3= the probability of low behavioral risk (Y=1). The simplistic multinomial logistic regression model relates the log of odds (or logit) of *Y* to the explanatory variable, *X*, in a linear form:

$$Logit(p_1) = natural\log(odds) = \ln(\frac{p_1}{1-p_1}) = a_1 + bX \qquad (1)$$

$$Logit(p + p_2) = natural\log(odds) = \ln_1(\frac{p_1 + p_2}{1-p_1-p_2}) = a_2 + bX. \qquad (2)$$

Note both equations (1) and (2) constitute one multinomial logistic model with the constraint that $\Sigma p_i = 1$. They model the cumulative probabilities with a common slope parameter (b) but different Y intercepts (α_1, α_2). The two *Y* intercepts are two constants in the multinomial logistic model; they are not a function of the predictor *X*.

The predictor, *X*, can be categorical or continuous while the outcome (*Y*) is always categorical. Parameters, α_1, α_2, and β, are typically estimated by the maximum likelihood (ML) method. The ML method is designed to maximize the likelihood of reproducing the data given their parameter estimates (Peng, Lee, et al., 2002). The value of the coefficient β reveals the direction of the relationship between *X* and the logit of *Y*. When β is greater than 0, larger (or smaller) *X* values are associated with larger (or smaller) logits of *Y*, and the curve will resemble an increasing sigmoid (or *S*-shape). Conversely, if β is less than 0, larger (or smaller) *X* values are associated with smaller (or larger) logits of *Y*. Such a relationship is often shown in data in the form of a reverse sigmoid curve. In other words, an increase in *X* is associated with a decrease in logits of *Y* and vice versa.

Within the framework of inferential statistics, the null hypothesis states that β equals zero in the population. Rejecting such a null hypothesis implies that a linear relationship exists between *X* and the logit of *Y*. If an explanatory variable is binary, such as gender in Table 1A and dropout in Table 1B, the β coefficient can also be interpreted as an odds ratio which numerically equals e (the natural logarithm base) raised to the exponent of β (i.e., e^{β}).

If two or more explanatory variables are included in the model (say X_1= gender and X_2= emotional risk score), one may construct a complex logistic regression for the logit of *Y* (high, medium, or low levels of behavioral risk) as follows:

$$Logit\ (p_1) = natural\ \ \log(odds) = \ln(\frac{p_1}{1-p_1}) = a_1 + b_1 X_2 + b_2 X_2 \qquad (3)$$

$$Logit(p_1) = natural\log(odds) = \ln(\frac{p_1 + p_2}{1-p_1-p_2}) = a_2 + b_1 X_1 + b_2 X_2. \qquad (4)$$

As noted before, equations (3) and (4) constitute one complex multinomial logistic model with the constraint that $\Sigma p_i = 1$. They model the cumulative probabilities with common slope parameters (ß$_1$ and ß$_2$) but different *Y* intercepts (α_1, α_2). The two *Y* intercepts are two constants in the multinomial logistic model; they are not a function of the explanatory variables. Explanatory variables, X_1 and X_2, can be categorical or continuous while the dependent variable (*Y*) is always categorical. Parameters, α_1, α_2, ß$_1$, and ß$_2$, are estimated by the maximum likelihood (ML) method, as in the simple multinomial model. Data are entered into the analysis as 1, 2, or 3 coding for the trichotomous dependent variable, continuous values for continuous explanatory variables, and dummy coding (e.g., 0 or 1) for categorical explanatory variables.

The null hypothesis underlying the complex multinomial logistic model states that all ß's equal zero. Rejecting this null hypothesis implies that at least one ß does not equal 0 in the population. The interpretation of ß is rendered using odds ratios. If ß$_j$ represents the regression coefficient for predictor X_j, exponentiating ß$_j$ yields the odds ratio (e^{ßj}). When all other explanatory variables are held at a constant, odds ratio is the change in the odds of *Y* given a unit change in X_j.

For the behavioral risk data, we hypothesized the following linear relationship might exist:

$$Logit(p_1) = \ln(\frac{p_1}{1-p_1}) = a_1 + b_1 X_2 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5, \qquad (5)$$

$$Logit(p_1 + p_2) = \ln(\frac{p_1 + p_2}{1-p_1-p_2}) = a_2 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5, \qquad (6)$$

where p_1= the probability of high behavioral risk (*Y*=3), p_2= the probability of medium behavioral risk (*Y*=2), and p_3= the probability of low behavioral risk (*Y*=1), X_1=GENDER (boys=1, girls=0), X_2=intention to drop out of school (DROPOUT, yes=1, no=0), X_3=family structure (FAMILY, intact family=1, step-family =2, and single-parent family=3), X_4=emotional risk score (EMOTION), and X_5=self-esteem score (ESTEEM).

Alternatively, one can express the same functional relationship by taking the antilog function of Equations (5) and (6) to obtain a direct estimate of the probabilities of behavioral risk:

$$p_1 = \frac{e^{a_1+b_1X_1+b_2X_2+b_3X_3+b_4X_4+b_5X_5}}{1+e^{a_1+b_1X_1+b_2X_2+b_3X_3+b_4X_4+b_5X_5}} \qquad (7)$$

$$p_1 + p_2 = \frac{e^{a_2+b_1X_1+b_2X_2+b_3X_3+b_4X_4+b_5X_5}}{1+e^{a_2+b_1X_1+b_2X_2+b_3X_3+b_4X_4+b_5X_5}} \qquad (8)$$

where e=2.71828 is the base of the system of natural logarithms. Equation (7) defines p_1 directly, whereas p_2 and p_3 are derived by subtraction; i.e., $p_2 = (p_1 + p_2) - p_1$ = equation 8 – equation 7, and $p_3 = 1 - (p_1 + p_2)$ = 1 – equation 8. As previously defined, p_1= the probability of high behavioral risk (Y=3), p_2= the probability of medium behavioral risk (Y=2), and p_3= the probability of low behavioral risk (Y=1).

Interpreting and Assessing Multinomial Logistic Regression Results

Equations (7) and (8) were fitted to data using SAS® PROC LOGISTIC (Version 8e, SAS Institute Inc., 1999) in order to support/refute the research hypothesis posed earlier that "the likelihood that an adolescent is at high, medium, or low behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self esteem." The result showed that

Predicted logit (Y1=High RISK)= -0.6211 + (1.1070)*GENDER + (2.1818)*DROPOUT + (0.4135)*FAMILY + (0.00738)*EMOTION + (-0.0488)*ESTEEM, (9)

and

Predicted logit (Y1+ Y2 =High + Medium RISK) = 2.5220 + (1.1070)*GENDER + (2.1818)*DROPOUT + (0.4135)*FAMILY + (0.00738)*EMOTION + (-0.0488)*ESTEEM (10)

The χ^2 test of proportional odds assumption was insignificant (df=5; p=0.6548), indicating that there was no need to fit a second model with distinct β parameters (Peterson & Harrell, 1990). Hence, Equations (9) and (10) will be hereafter referred to as the MLR model.

Interpreting Multinomial Logistic Regression Results

According to the MLR model, the log of the odds of an adolescent's behavioral risk level was positively related to gender (p<.0001, Table 2), intention to drop out of school (p<.0001), and family structure (p<.001); it was negatively related to self-esteem (p<.0001), and insignificantly related to emotional risk (p =0.5211). The positive coefficient (1.1070) associated with GENDER in the MLR model implied that boys were more likely, than girls, to be at high behavioral risk, holding all other explanatory variables constant. In fact, the odds of a boy being at high behavioral risk were 3.025 (= $e^{1.1070}$, Table 2) times greater than the odds for a girl. The same trend was observed with the dichotomous variable of DROPOUT from school. The odds of teen-age students engaging in high or medium risk of behavior, than not, were 8.8622 times higher for students intending to drop out than students without such an intention. This relationship can also be seen in Table 1B in which the majority of those intending to not stay in school were placed in high or medium level of behavioral risk, compared to those with intentions to stay in school.

Regarding the third categorical variable family structure, interpretation should be based on the reference group of intact families. Thus, the higher the score on FAMILY, the less stability in the family structure and the greater is the behavioral risk for adolescents. This interpretation was rendered by the positive coefficient associated with FAMILY. As a family's structure changed from 1 (intact family) to 2 (step family) or from 2 to 3 (single family), the odds increased by 1.5121 for adolescents to be at a higher behavioral risk than medium or low risk.

The coefficient for self-ESTEEM indicated that the decrease in log odds of risky behavior corresponded to one unit increase in self-esteem scores. In other words, the higher the self-esteem score, the less likely an adolescent would be at high behavioral risk. For each point increase on the self esteem score, the odds of participating in risky behavior, compared to the odds of not participating, decreased from one to 0.952 (= $e^{-0.0488}$, Table 2). If the increase on the self-esteem score was 10 points, the odds decreased from one to 0.6139 [= $e^{10*(-0.0488)}$].

Combining the four explanatory variables that were found to be statistically significant in the MLR model, a profile emerged for a youth at the greatest predicted behavioral risk: a male who intended to drop out of school, came from a single parent household, scored low on the self-esteem measure, and possibly high on the emotional risk measure (based on the positive correlation between behavioral risk and emotional risk) — this last characteristic did not reach statistical significance in the MLR model.

Assessing Multinomial Logistic Regression Results

How effective was the MLR model expressed in Equations (9) and (10)? How can a health educator assess the soundness of a multinomial model? To answer these questions, we attended to (a) overall model evaluations, (b) statistical tests of each explanatory variable, (c) goodness-of-fit statistics, and (d) validations of predicted probabilities. These evaluations are discussed below based on Equations (9) and (10), or the MLR model.

(a) Overall model evaluations. The Likelihood Ratio, Score, and Wald tests were examined to determine the improvement of the MLR model over the intercept-only model (also called the null model). According to Peng, Lee, and Ingersoll (2002, p.6), “An intercept-only model serves as a good baseline because it contains no predictors; consequently all observations would be predicted to belong in the largest outcome category, according to this model.” All three tests yielded similar results (p<.0001, Table 2), namely, the MLR Model was more effective than the null model. It was therefore inferred that at least one explanatory variable was a significant predictor of adolescents’ behavioral risk. After splitting the sample randomly 5 times, resulting in 10 random sub-samples, we applied the same multinomial model to the sub-samples. The overall significance of the MLR model was replicated in all 10 sub-samples.

(b) Statistical tests of individual predictors. The individual β coefficients were tested using the Wald χ^2 statistic (Table 2). All variables except for EMOTION were significant predictors of adolescents’ risk for self-injurious behaviors (p<.001). Two predictors (GENDER, and ESTEEM) were cross-validated to be significant; one predictor (EMOTION) was replicated to be statistically insignificant, all with 10 random sub-samples. FAMILY structure and intention to DROPOUT were confirmed to be statistically significant predictors in 9 out of 10 cross-validation random samples. It was not necessary to statistically test the intercepts for the two constants (CONSTANTs 1 and 2 in Table 2) as the test result merely indicates if intercepts should be included in a logistic model (Peng, Lee, & Ingersoll, 2002).

(c) Goodness-of-fit statistics. Goodness-of-fit statistics assess the fit of a logistic model against actual classifications, i.e., high, medium, or low level of behavioral risk. Two descriptive measures of goodness-of fit are presented in Table 2 for the MLR model: R^2 indices defined by Cox and Snell (1989) and Nagelkerke (1991), respectively. These two measures were similar for the MLR model (24.67% and 29.78%). According to Peng, Lee, and Ingersoll (2002), these indices are variations of the R^2 concept defined for the ordinary least squares (OLS) regression model.

Even though the R^2 has a clear definition in OLS regression, there have been no equivalents of this concept devised by methodologists for multinomial logistic models that render the meaning of variance explained; none correspond to predictive efficiency, and none can be tested in an inferential framework (Mendard, 2000). For these reasons, a researcher may treat these two R^2 indices reported in Table 2 as supplementary to other, more useful evaluative indices such as the overall evaluation of the model, tests of individual regression coefficients, and the inferential test of the goodness-of-fit suggested by Begg and Gray (1984) for multinomial logistic models.

Table 2. Multinomial Logistic Regression Analysis of Adolescent's Self-inflicting Behavior Risk by SAS® PROC LOGISTIC (version 8).

Predictor	*b*	*SE b*	Wald's χ^2 (*df*=1)	*p*	e^b *(odds ratio)*
CONSTANT 1 (Y_1)	–0.6211	1.0627	0.3416	0.5589	Not necessary
CONSTANT 2 (Y_1+Y_2)	2.5220	1.0723	5.5317	0.0187	Not necessary
GENDER (boys=1,girls=0)	1.1070	0.2111	27.5060	<0.0001	3.0253
DROPOUT (yes=1, no=0)	2.1818	0.3287	44.0618	<0.0001	8.8622
FAMILY	0.4135	0.1179	12.2979	<0.001	1.5121
EMOTION	0.0074	0.0115	0.4118	0.5211	1.0074
ESTEEM	–0.0488	0.0118	16.9867	<0.0001	0.9524

Overall Model Evaluation

Tests	χ^2	*df*	*p*
Likelihood Ratio Test	122.38	5	<0.0001
Score test	110.47	5	<0.0001
Wald test	97.87	5	<0.0001

Notes. Cox and Snell *R* squared=0.2467. Nagelkerke *R* squared (Max rescaled *R* squared)=0.2978. Kendall's Tau-*a* = 0.297. Goodman-Kruskal's Gamma= 0.548. Somers' D_{xy}= 0.539. *c*-statistic = 0.769.

SAS® Programming Codes

```
PROC LOGISTIC DATA=risk432
        MODEL risk= gender dropout family emotion esteem;
        OUTPUT out=probs  predicted=prob xbeta=logit;
RUN;
```

According to Begg and Gray (1984, cited in Hosmer & Lemeshow, 2001, p. 281), the goodness-of-fit test of a multinomial model may be carried out by applying the Hosmer and Lemeshow (H-L) test to two of the three outcome categories, then integrating the test results descriptively. For the logistic model comparing low risk adolescents with medium risk adolescents, the H-L test yielded a χ^2 of 5.8011 with 8 degrees of freedom. For the logistic model comparing low risk adolescents with high risk adolescents, the H-L test yielded a χ^2 of 8.2925, also with 8 degrees of freedom. Both test results were statistically insignificant (p>.40) indicating that both models fit the data well. In other words, the null hypothesis of a good model fit to data was tenable.

(d) Validations of predicted probabilities. As was explained previously, the MLR model predicts the logit of high and medium levels of behavioral risk from a set of explanatory variables. Since logit is the natural log of the odds [or probability/ (1-probability)], it can be transformed back to the probability scale, according to Equations (7) and (8). Once the predicted probability of behavioral risk is calculated, it can be compared with the actual risk behavior to determine if high probabilities are associated with the high level of behavioral risk, low probabilities with the low level of behavioral risk, and middle-range probabilities with the medium level of behavioral risk.

SAS® PROC LOGISITC (version 8) provides four measures of association for logistic regression models. These are: Kendall's Tau-*a*, Goodman-Kruskal's Gamma, Somers' *D* statistic, and the *c* statistic (Table 2). Kendall's Tau-*a* is a rank-order correlation coefficient without adjustments for ties; for the MLR model, it equaled 0.287. Goodman-Kruskal's Gamma equaled 0.548. According to Peng, Lee, and Ingersoll (2002), it is a more useful and appropriate measure than Tau-*a* when there are ties on both dependent variable categories and predicted probabilities (the present data had 923 ties — approximately 1.8% of all pairs). This measure is interpreted as 54.8% fewer errors made in predicting which of two adolescents would be at a greater behavioral risk by utilizing the estimated probabilities, than by chance alone (Demaris, 1992). Some caution is advised in using the Gamma statistic since: (1) it has a tendency to overstate the strength of association between estimated probabilities and outcomes (Demaris), and (2) a value of zero does not necessarily imply independence when the data structure exceeds a 2 by 2 format (Siegel & Castellan, 1988).

Somers' *D* is a preferred extension of Gamma whereby one variable is designated as the dependent variable and the other the independent variable (Siegel & Castellan, 1988). For the MLR model, Somers' *D* was 0.539 (Table 2). There are two asymmetric forms of Somers' *D* statistic: D_{xy} and D_{yx}. Only D_{yx} correctly represents the degree of association between the behavioral risk level (*y*), designated as the dependent variable, and the estimated probability (*x*), designated as the independent variable (Demaris, 1992).

Unfortunately, SAS® computes only D_{xy}, although this index can be corrected to D_{yx} in SAS® (Peng & So, 1998). For the present model, the *c* statistic was 0.769 (Table 2). This means that for 76.90% of all possible pairs of adolescents, one at a greater risk (e.g., high or medium level) than the other (e.g., medium or low level), the MLR model correctly assigned a higher probability to those measured by HBQ at greater behavioral risk. Thus the model worked better than assigning observations randomly into categories of high, medium, or low behavior risk. The *c* statistic ranges from 0.5 to 1.

A 0.5 value means that the model is no better than assigning observations randomly into categories of the dependent variable. A value of 1 means that the assignment of probabilities matches perfectly with the ordered categories of the dependent variable (e.g., high with high, medium with medium, and low with low). If several models were fitted to the same data, the model chosen as the "best" model should be associated with the highest *c* statistic. Thus, the *c* statistic provides a basis for comparing different models fitted to the same data, or the same model fitted to different data sets.

Reporting Multinomial Logistic Regression Results

In addition to Tables 1 and 2, it is helpful to profile adolescents with certain characteristics and relate these characteristics to the predicted probability of engaging in high, medium, or low level of risky behaviors. For this purpose, several boys and girls, from either an intact, step-parent, or single-parent home, were selected from the data base. These characteristics, along with their indication to stay in or drop out of school and their emotional risk and self-esteem measure, are shown in Table 3 (following References section) to be related to their predicted probability of engaging in various levels of risky behaviors. It is noted in Table 3 that 8 cases (#6, 12, 19, 22, 30, 31, 34, and 36) did not exist in the data. These cases may be explained by their refusal to participate, missing data (to be addressed in the next section), and the improbable likelihood of locating these adolescents in the population (e.g., case #30, 31, 34, and 36).

Among boys from the intact family (cases #1 to #5), the probability of engaging in low-level of risky behaviors (#3) was associated with a very low emotional risk score and no intention to drop out of school. Likewise, girls from the intact family (cases #7 to #11), who were predicted to engage in low-level of risky behaviors, did not intend to drop out from school and were measured low on emotional risk.

Boys from the step-parent family (#13 to #18), were predicted to engage in medium to high level of risky behaviors. The higher the emotional risk score, the greater was the probability of being associated with high-risk behaviors (#18). For girls from step-parent families (#20, 21, 23, and 24), those with no intention to drop out of school (#20 and #21) were predicted to engage in lower levels of risky behaviors than those with an intention to drop out of school.

Among boys from the single-parent home (#25 to #29), engaging in high-level risky behaviors was predicted for the boy with an intention to drop out of school (#29), whereas low-level was predicted for the boy who had no intention to drop out of school, scored low on the emotion risk measure, and high on the self-esteem test (#26). Among girls from the single-parent home (#32, 33, and 35), all were predicted to engage in medium level of risky behaviors. Though cases #32 and #33 did not intend to drop out of school, they scored high on emotional risk and low on self-esteem. Case #35 intended to drop out of school; she was measured comparatively low on emotional risk and high on self-esteem.

Missing Data

It is important to point out the problem with missing data encountered in the multinomial logistic modeling, especially for the explanatory variable emotional risk (EMOTION). Descriptive analyses of the data suggested one plausible explanation for the insignificant relationship between emotional risk and behavioral risk (Table 2). Of the 85 cases with missing data, 77 were missing behavioral risk data, 34 were missing emotional risk data, and six were missing drop-out scores. It was noted that the range (34.21 to 82.03), mean (50.11), and standard deviation (10.94) for the 51 (=85–34) emotional risk scores not included in the analysis, were slightly higher than those used in the analysis. Furthermore, 25 (or 49.02%) of the 51 emotional risk scores were above the overall sample mean of 48.72. It would be important to ascertain why adolescents with slightly higher emotional risk scores chose not to complete the behavioral risk assessment. Thus, missing data on the dependent variable might not be missing completely at random (Little and Rubin, 1987).

To answer this question statistically, we imputed all missing values using the EM method installed in the MVA (missing value analysis) module of SPSS Version 11.01. The complete data set with imputed values (N=517 observations) contained 255 adolescents at low behavioral risk, 228 at medium risk, and 34 at high risk. The complete data set was submitted to SAS® PROC LOGISTIC (Version 8e) for multinomial logistic regression modeling. Results were very similar to those in Table 2, namely, gender, intention to drop out from school, family structure, and self-esteem were statistically significant at p<.0001. The emotional risk variable was again not a statistically significant predictor. An examination of correlations between the behavioral risk level and the five predictors showed that the positive correlation between emotional risk scores and the behavioral risk level, though positive, was not as high as the correlation between self-esteem scores and behavioral risks. And there was a strong negative correlation between emotion risk and self-esteem (Pearson r = -.494). Based on these results, we concluded that the missing data did not bias the interpretations given earlier for the MLR model.

Conclusion

In this article, we applied multinomial logistic regression to data based on 432 adolescents' self-reported measures of behavioral risk, emotional risk, self-esteem, intention to drop out of school, and their gender and family structure to test a research hypothesis. The research hypothesis stated that, "the likelihood that an adolescent child is at high, medium, or low level of self-injurious behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self esteem." Logistic regression results supported the statistical significance of four explanatory variables.

Specifically, the likelihood of an adolescent participating in risky behaviors was negatively related to his/her self-esteem scores, but positively related to intention to drop out of school, family structure, and gender. If all other explanatory variables were held as constants, adolescents with the following profiles were more likely, than their counterparts, to engage in risky behaviors: boys, intending to drop out of school, living in a single-parent household, and having low self-esteem. The effectiveness of the multinomial logistic model was supported by multiple indices, including the model's overall test of all explanatory variables, statistical significance test of each explanatory variable, the predictive power of the model, and its interpretability.

Three methodological issues encountered during the logistic regression analysis were highlighted and treated in our discussion of the results. These included (1) the use of odds ratio in interpreting results obtained from MLR models, (2) the absence of an extension of the Hosmer and Lemeshow goodness-of-fit test for multinomial logistic models, and (3) the missing data problem.

From the standpoint of modeling categorical outcomes, logistic regression is more flexible and less restrictive than discriminant function analysis, log-linear models, or modified probability models (Peng, Manz, & Keck, 2001). While logistic regression is gaining popularity in health and social sciences research (Peng, Lee, & Ingersoll, 2002; Peng, So, Stage, & St. John, 2002), there are few studies that demonstrate a preferred pattern of the application of multinomial logistic regression methods. It is hoped that this paper has demonstrated that multinomial logistic regression is an effective technique for profiling those youth at greatest risk for participation in risky health behaviors. Psychologists and educators can utilize findings to plan prevention programs, as well as to apply the versatile logistic technique in psychological, educational, and health research concerning adolescents.

References

Austin, J. T., Yaffee, R. A., & Hinkle, D. E. (1992). Logistic regression for research in higher education. *Higher education: Handbook of theory and research, Vol. VIII,* 379-410.

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

Brack, C. J., Orr, D. P., & Ingersoll, G. (1988). Pubertal maturationand adolescent self-esteem. *Journal of Adolescent Health Care, 9,* 280-285.

Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. *Higher education: Handbook of theory and research, Vol. X,* 225-256.

Chuang, H. L. (1997). High school youth's dropout and re-enrollment behavior. *Economics of Education Review*, *16*(2), 171-186.

Demaris, A. (1992). *Logit modeling: Practical applications*. Newbury Park, CA: Sage.

Everett, S. & Husten, C. G. (1999). Smoking initiation and smoking patterns among US college students. *Journal of American College Health, 48*(2), 55-61.

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

Ingersoll, G. M. & Orr, D. P. (1989). Behavioral and emotional risk in early adolescents. *Journal of Early Adolescence*, *9*, 392-408.

Ingersoll, G. M., Grizzle, K., Beiter, M. & Orr, D. P. (1993). Frequent somatic complaints and psychosocial risk in adolescents. *Journal of Early Adolescence, 13*(1), 67-78.

Janik, J., & Kravitz, H. M. (1994). Linking work and domestic problem with police suicide. *Suicide and Life Threatening Behavior, 24*(3), 267-274.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17-24.

Morgan, S. P, & Teachman, J.D. (1988). Logistic regression: Description, examples, and comparisons. *Journal of Marriage and the Family, 50*, 929-936.

Nunnally, J. (1977). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Orr, D. P., Wilbrandt, M. L., Brack, C. J., Rauch, S. P., & Ingersoll, G. M. (1989). Reported sexual behaviors and self-esteem among young adolescents. *American Journal of Diseases of Children, 143,* 86-90.

Peng, C. Y., & So, T. S. (1998). If there is a will, there is a way: Getting around defaults of PROC LOGISTIC in SAS. *Proceedings of the MidWest SAS Users Group 1998 Conference,* 243-252. (php.indiana.edu/~tso/articles/mwsug98.pdf)

Peng, C. Y., Manz, B. D., & Keck, J. (2001). Modeling categorical variables by logistic regression. *American Journal of Health Behavior, 25*(3), 278-284.

Peng, C. Y., Lee, K. L., & Ingersoll, G.M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*(1), 3-14.

Peng, C. Y., So, T. S. H., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education, 43(3)*, 259-293.

Peterson, B. & Harrell, F. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics, 39*, 205-217.

Resnick, M. D., Harris, L. J., & Blum R.B. (1993). The impact of caring and connectedness on adolescent health and well-being. *Journal of Pediatrics Child Health, 29*(1), 53-59.

Robins, P. K., & Dickinson, K. P. (1985). Child support and welfare dependence: A multinomial logit analysis. *Demography, 22(*3), 367-380.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rouse, K. A. G., Ingersoll, G. M., & Orr, D. P. (1998). Longitudinal health endangering behavior risk among resilient and nonresilient early adolescents. *Journal of Adolescent Health, 23*, 297-302.

SAS Institute Inc. (1995). *Logistic regression examples: Using the SAS®, version 6, first ed.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1999). *SAS/STAT® user's guide, version 8, volume 2.* Cary, NC: SAS Institute Inc.

Scott, K. G., Mason, C. A., & Chapman, D. A. (1999). The use of epidemiological methodology as a means of influencing public policy. *Child Development, 70(5)*, 1263-1272.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral science* (2nd ed.). New York: McGraw-Hill.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Tolman, R.M., & Weisz, A. (1995). Coordinated community intervention for domestic violence: the effects of arrest and prosecution on recidivism of woman abuse perpetrators. *Crime and Delinquency, 41*(4), 481-495.

Yarandi, H. N., & Simpson, S. H. (1991). The logistic regression model and the odds of testing HIV positive. *Nursing Research, 40*(6), 372-373.

Table 3. Predicated Probability of Participating in Self-injurious Behavior for 36 Children.

Case	SEX	DROPOUT	FAMILY	EMOTION	ESTEEM	Intercept	Intercept	Predicted probability			Actual Behavior risk,
No.	ß=	ß=2.1818	ß=0.4135	ß=0.0074	ß=	1	2	of participating in			1=high, 2=med, 3=low
	1.107	1=yes	1=intact,		−0.0488	α_1 =	α_2 =2.522	self-injurious behavior			(score on HBO,
	1=boy	0=no	2=step,			−0.6211		p_1	p_2	p_3	M=47.69, SD=10.89)
	0=girl		3=single					(high)	(medium)	(low)	
1	1	0	1	62.39	32.68	−0.6211	2.5220	.0818	.5921	.3261	1 (60.40)
2	1	0	1	80.74	32.68	−0.6211	2.5220	.0926	.6102	.2972	2 (52.77)
3	1	0	1	32.07	71.58	−0.6211	2.5220	.0106	.1878	.8016	3 (42.65)
4	1	1	1	72.72	46.41	−0.6211	2.5220	.3038	.6062	.0900	1 (95.21)
5	1	1	1	63.07	37.25	−0.6211	2.5220	.3885	.5479	.0636	2 (50.00)
6	1	1	1	----	----	−0.6211	2.5220	---	---	---	3 (------)
7	0	0	1	47.29	41.83	−0.6211	2.5220	.0166	.2645	.7189	1 (61.53)
8	0	0	1	45.78	44.12	−0.6211	2.5220	.0147	.2422	.7431	2 (47.07)
9	0	0	1	42.05	21.24	−0.6211	2.5220	.0425	.4643	.4932	3 (42.70)
10	0	1	1	51.37	34.97	−0.6211	2.5220	.1772	.6559	.1669	1 (70.23)
11	0	1	1	56.77	37.25	−0.6211	2.5220	.1670	.6559	.1771	2 (53.27)
12	0	1	1	----	----	−0.6211	2.5220	---	---	---	3 (-------)
13	1	0	2	41.36	50.98	−0.6211	2.5220	.0451	.4776	.4773	1 (72.83)
14	1	0	2	46.14	50.98	−0.6211	2.5220	.0467	.4848	.4685	2 (45.84)
15	1	0	2	36.11	41.83	−0.6211	2.5220	.0663	.5559	.3778	3 (40.44)
16	1	1	2	38.59	57.85	−0.6211	2.5220	.2269	.6449	.1282	1 (92.50)
17	1	1	2	54.87	46.41	−0.6211	2.5220	.3665	.5641	.0694	2 (46.99)
18	1	1	2	70.35	34.97	−0.6211	2.5220	.5312	.4321	.0367	3 (43.52)
19	0	0	2	---	---	−0.6211	2.5220	---	---	---	1 (-------)
20	0	0	2	34.21	44.12	−0.6211	2.5220	.0203	.3041	.6756	2 (45.78)
21	0	0	2	50.18	53.27	−0.6211	2.5220	.0147	.2421	.7432	3 (40.44)
22	0	1	2	----	---	−0.6211	2.5220	---	---	----	1 (------)
23	0	1	2	54.84	50.98	−0.6211	2.5220	.1326	.6473	.2201	2 (48.64)
24	0	1	2	50.18	46.41	−0.6211	2.5220	.1559	.6547	.1894	3 (43.08)
25	1	0	3	63.52	23.52	−0.6211	2.5220	.2432	.6384	.1184	1 (67.90)
26	1	0	3	32.07	67.00	−0.6211	2.5220	.0296	.3848	.5856	2 (56.69)
27	1	0	3	50.18	48.70	−0.6211	2.5220	.0786	.5854	.3360	3 (40.44)
28	1	1	3	43.54	48.70	−0.6211	2.5220	.4184	.5250	.0566	1 (85.49)
29	1	1	3	56.74	44.12	−0.6211	2.5220	.4979	.4604	.0417	2 (54.31)
30	1	1	3	---	---	−0.6211	2.5220	---	---	----	3 (-------)
31	0	0	3	---	---	−0.6211	2.5220	---	---	---	1 (-------)
32	0	0	3	64.12	28.10	−0.6211	2.5220	.0786	.5856	.3358	2 (48.41)
33	0	0	3	60.08	39.54	−0.6211	2.5220	.0453	.4781	.4766	3 (44.41)
34	0	1	3	---	---	−0.6211	2.5220	---	---	---	1 (-------)
35	0	1	3	43.63	48.70	−0.6211	2.5220	.1922	.6543	.1535	2 (46.34)
36	0	1	3	---	---	−0.6211	2.5220	---	---	----	3 (-------)