Logit Regression: Best Model Selection

Randall E. Schumacker, University of North Texas
Cynthia Anderson, University of North Texas
James Ashby, Richardson ISD, Texas

The typical method of analyzing categorical variables is to use the chi-square statistic. However, with more than two categorical variables, simultaneous examination of main and interaction effects is not feasible. The logit regression technique permits analysis of categorical variables, the modeling of main and interaction effects, control of Type I error, and distribution freer assumptions. This study investigated parsimonious model fit related to the selection of the best set of categorical predictor variables. Findings indicated that the various variable selection criteria (L^2, z, log-odds ratio, R^2_L, model variance, and ΔC^2) provided different results. Order of variable entry also produced significantly different results. The use of a Tabu search procedure and ΔC^2 criteria is recommended to determine the best set of categorical independent predictor variables in logit regression.

Logit regression is a special case of log linear regression where both the dependent and independent variables are categorical in nature (Klienbaum, 1992). It offers distinct advantages over the chi-square method for analysis of categorical variables. Some of these advantages are: (1) control of Type I error rates, (2) modeling of interaction effects, and (3) distribution freer assumptions. The main objective of this study was to investigate the selection of the best set of categorical predictor variables in the presence of main and interaction effects. In logit modeling, natural log odds of the frequencies are computed which allow different models and different model parameters to be compared given the additive nature of the L^2 component for each model.

Logit regression is affected by sample size, outliers and inadequate expected frequencies in categorical cells (Demaris, 1992). This often occurs with too many categorical variables and small sample sizes, hence inadequate cell sizes. In has been understood that cell size should not have fewer than $n = 5$ (Hinkle, McLaughlin, & Austin, 1998; Kennedy, 1992). Another rule of thumb indicated that total sample size should be at least 4 to 5 times the number of cells in the model (Feinberg, 1981). Marasculio and Busk (1987) suggested that low expectancy in cells, possibly due to rare events, should be sampled until adequately filled, and if outliers are suspected, residuals be examined. Collapsing categories is also a reasonable option.

A theoretical logit regression model is generally postulated (null model or base model). A common practice is then to create one or more hierarchical models where each new model contains parameters of the previous model, plus a hypothesized new parameter. The theoretical model can be tested beginning with a null model and adding parameters, or with a saturated model deleting parameters. The best model is selected based on the likelihood ratio statistic, L^2. If the likelihood ratio statistic is significant, then the observed frequencies do not fit the expected frequencies, or in other terms, the data doesn't fit the theoretical model (hypothesized logit regression equation). Several logit regression models may "fit" equally well. In this case, the non-significant likelihood ratio statistics' for the competing models are subtracted yielding a L^2 difference test of model fit analogous to the change in R^2 in regression analysis. If the model change is not significant, then the most parsimonious model is typically chosen. Identification of significant variable parameters in the model is assessed by partitioning the L^2 into its additive components relative to the specified model. Post-hoc procedures generally evaluate fit of the data to individual cells based on standardized residuals or variance accounted for in the model.

Various criteria can be used to determine the predictors to include in a logit regression model:

1. Pearson chi-square or likelihood-ratio χ^2
2. z-test of parameters in model
3. log-odds ratio
4. Predictive efficiency (R^2 type measure)
5. ΔC^2 (difference between -2logL values for null and model)

The traditional Pearson chi-square and the likelihood-ratio chi-square with $(I\text{-}1)(J\text{-}1)$ degrees of freedom are similar because, as sample size becomes larger, the sampling distributions of both statistics become asymptotically chi-squared. The likelihood-ratio chi-square is computed as:

$$L^2 = 2\Sigma\Sigma n_{ij} \log[n_{ij}/ m_{ij}];$$

where n_{ij} = observed cell frequency, m_{ij} = expected cell frequency (Demaris, 1992, *p.* 4)..

The parameter estimates calculated using maximum likelihood estimation possess asymptotic properties. As sample size increases, the parameter

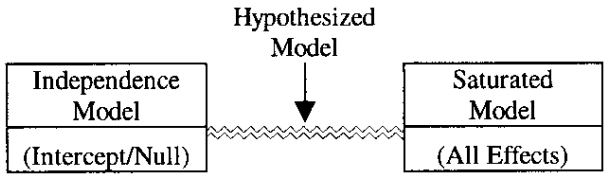
estimates become unbiased and consistent with population parameters. The sampling distribution also approaches normality with variance lower than other unbiased estimation procedures (least squares, etc.). Therefore, given larger samples, the test of a parameters'significance (independent categorical predictor variable) is a z-test calculated by:

$$Z = \hat{\beta}^1 / SE(\hat{\beta}^1)$$

Parameter estimates in logit models can also be readily interpreted as a log-odds ratio. This is calculated as e^{β} for a single parameter, or $e^{(\beta 1 - \beta 2)}$ for differences between two parameters. This is useful when examining contrasts between levels of two independent categorical predictor variables. The log-odds ratio will always agree with the expected cell frequencies.

Predictive efficacy refers to whether a model generates accurate predictions of group membership on the dependent variable. It is possible to have an excellent fit between the logit model and the data without having predictive efficacy. Recall, if $L^2 = 0$, a saturated model exists which perfectly fits the data, yet predictive efficacy (classification) can be far from perfect. In ordinary least squares regression, a saturated model would yield $R^2 = 1$. In SPSS, a classification coefficient (*c*) is calculated to indicate the amount of variance in the dependent variable accounted for by a set of predictors in the model. Given a 50-50 sample split, the base percent would be 50% for the independence model (intercept only model), thus *c* for each variable should be interpreted as a percent that contains this base percentage.

The R^2_L - type measure for logistic regression is not meant as a variance accounted for interpretation, as traditionally noted in least squares regression, because it under estimates the proportion of variance explained in the underlying continuous variables (an assumption made about categorical variables). Basically, a loss of power results when data are reduced from interval to ordinal to nominal. Instead, the R^2_L - type measure is an approximation (lower bound) for assessing predictive efficacy ranging from zero (0) [independence model] to one (1) [saturated model]. This can be depicted as:



The R^2_L-type measure (Hosmer & Lemeshow, 1989) is calculated as: $R^2_L = (SS_T - SS_E)/ SS_T$, where $SS_T = -2\log L_0$ and $SS_E = -2\log L_1$.

The ΔC^2 value provides a way to examine alternative logit models. The L^2 from one model is simply subtracted from the L^2 of the second model. This is similar to testing a full versus restricted modl in multiple regression. The calculation is simply: $L^2(2|1) = L^2_2 - L^2_1$ with the degrees of freedom equal to the difference in the degrees of freedom of the two models. In terms of the log values it is

$$C^2 = -2\log L_0 - (-2\log L_1).$$

If C^2 is non-significant, then additional independent categorical predictors in Model 2 are not needed. This type of test is only appropriate for the likelihood-ratio chi-square and not the Pearson chi-square because adding additional independent categorical predictor variables will never result in a poorer fit of the model to the data (similar to adding terms to a regression model that will never yield a lower unadjusted R^2). This property doesn't hold for the Pearson chi-square.

Logit Models

The logit model contains a categorical dependent variable and a set of categorical independent predictor variables. If a non-significant likelihood-ratio chi-square (L^2) value is computed, then a given model fits the observed data, which is what we desire. On one extreme of the logit model continuum is the *saturated model* or model with perfect fit, yielding a $L^2 = 0$ and *df*=0. The saturated model has as many parameter estimates as degrees of freedom, so it always perfectly reproduces the cell frequencies. For example, a model with all variable main effects and all interaction effects would lead to a saturated model. The *independence model*, in contrast, sets all parameter estimates to zero, resulting in the null model or intercept only model. Consequently, we have a model continuum ranging from the saturated model (all parameters estimated) to the independence model (no parameters estimated). A *hypothesized model* should fall somewhere between these two end-points and reflect a model with fewer parameter estimates than degrees of freedom, so that the degrees of freedom equals the total number of cells minus the number of parameters to be estimated in the model. A model containing only main effects would be an example.

The problem for a researcher becomes one of finding the best set of independent categorical predictor variables. However, what criteria should a researcher use to determine data-to-model fit? Oftentimes, main effects and/or interaction effects are included in a model to predict a dependent variable. For the purpose of this study, two examples are given which focus on the prediction of high-school dropout percent given a set of independent categorical predictor variables. Can we

Table 1. Logit Regression Models

	Loglinear Model	Model Designation
1	$\lambda_i + \lambda_j + \lambda_k + \lambda^R_{ij}$	[RL, D]
2	$\lambda_i + \lambda_j + \lambda_k + \lambda^R_{ij} + \lambda^R_{ik}$	[RL, RD]
3	$\lambda_i + \lambda_j + \lambda_k + \lambda^R_{ij} + \lambda^R_{ik} + \lambda^L_{jk}$	[RL, RD, LD]
4	$\lambda_i + \lambda_j + \lambda_k + \lambda^R_{ij} + \lambda^R_{ik} + \lambda^L_{jk} + \lambda^R_{ijk}$	[RLD]

predict dropout/non-dropout status based on a set of independent categorical predictor variables? Given this research question, we were concerned with the predictive efficacy of the logit model.

Study One

Method and Data

The National Education Longitudinal study of 1988 (NELS) data base was used for data analysis. Subjects were 391 twelfth grade students selected randomly from the NELS data base. Dropout status was treated as the categorical dependent variable. Grade repeat status and locus of control were designated as categorical independent variables. The main and interaction effects research questions were: (1) Do drop-out rates differ significantly between students who repeat a grade versus not repeat a grade?; (2) Do drop-out rates differ significantly between students who have high versus low locus of control?; and (3) Do drop-out rates differ significantly given an interaction between grade repeat status and locus of control? This basic study analysis was gleamed from a previous presentation by Anderson (1995).

The null model and alternative models are specified in Table 1. Model 1 is a null model which hypothesized that drop-out rates (D) are the same regardless of grade repeat status (R) and locus of control (L). Model 2 hypothesized a main effect for grade repeat status (R). Model 3 hypothesized a main effect for locus of control (L). Model 4 hypothesized an interaction between grade repeat status (R) and locus of control (L) in predicting drop-out rates (D).

Results

The calculation of L^2 is affected by the order of entry of independent categorical variables, consequently Table 2 indicates grade repeat status entered first (Method A) compared to locus of control entered first (Method B). Method A indicated a non-significant main effects L^2 value for grade repeat status and locus of control. Method B however indicated a non-significant main effects L^2 value for grade repeat status only. No interaction was indicated. A subsequent approach was to use the additive properties of the likelihood ratio statistic to assess the specific contribution of each parameter in the model specified by calculating the L^2 difference. Table 3 indicates the component L^2 values which are the difference between two modeled L^2 values. Model 2 (grade repeat status main effect) is statistically significant accounting for 93% of the total modeled L^2. Locus of control main effect and interaction effects are not significant.

The variance accounted for approach is yet another way to assess how much of the Null Model L^2 (48.58) is attributed to a hypothesized logit model, in this case Model 2 L^2 (45.13) in Table 2. It follows that 45.13 divided by 48.58 equals 93% of the Total L^2. Obviously, the other modeled L^2 values account for the remaining percent of the Total L^2. SPSS does compute and list a *c* value which indicates the percent classification.

Several post-hoc procedures have been suggested including standardized residuals (Hinkle et al., 1988), scheffe-type contrasts (Marascuilo & Busk, 1987), log-odds ratio of parameter estimates (Kennedy, 1992), and variance accounted for indicated above. A further investigation of this technique and analysis is presented in a second study to clarify best model selection strategies given multiple categorical independent predictor variables in logit regression models.

Study Two

Method and Data

There were 29,124 students enrolled in grades 7-12 in Richardson ISD. Of these students, 754 were dropouts (2.6%) and 28,370 were non-dropouts (97.4%). To facilitate the analysis, a random sample of 754 students was taken from the non-dropout students. The dependent variable was dropout status (dropout, non-dropout). The categorical independent predictor variables were: gender (male, female); ethnicity (asian, black, hispanic, white); grade (7,8,9,10,11,12); retained in grade (not retained, retained 1+ times); parent (natural, step/in-law); suspensions from school (none, 1, 2+); economic disadvantaged (no, yes); and number of courses missed (none, 1-5, 6+).

The research question of interest was in predicting dropout/non-dropout status from several independent categorical predictor variables. Consequently, predictive efficacy or classification status was the focus of the study. Basically, What set of independent predictors provides the best classification of dropout/non-dropout?

Table 2. Logit Regression Model Fit.

Method A	Residual			Component			
Model	L^2	*df*	*p*		L^2	*df*	*p*
Null Model – (1) [RL, D]	48.58	3	.0001				
Main Effects – (2) Grade Repeat [RL, RD]	3.45	2	.1779	$L^2_{(1-2)}$	45.13	1	.0001
Main Effects – (3) Locus of Control [RL, RD, LD]	0.24	1	.6242	$L^2_{(2-3)}$	3.21	1	.0421
Interaction – (4) Grade Repeat x Locus of Control [RLD]	0	0	---	$L^2_{(3-4)}$	0.24	1	.6242
Method B							
Null Model – (1) [RL, D]	48.58	3	.0001				
Main Effects – (2) Locus of Control [LR, LD]	41.85	2	.0001	$L^2_{(1-2)}$	6.73	1	.0071
Main Effects – (3) Grade Repeat [LR, LD, RD]	0.24	1	.6242	$L^2_{(2-3)}$	41.61	1	.0001
Interaction – (4) Grade Repeat x Locus of Control [RLD]	0	0	---	$L^2_{(3-4)}$	0.24	1	.6242

Results

A preliminary univariate analysis of each categorical independent predictor with the dependent variable dropout status is in Table 3. It is apparent that gender differences are not significant in determining dropout/non-dropout status. Similarly, economic disadvantaged doesn't yield a high L^2 or χ^2 relative to the other predictor variables. The slight difference in L^2 and χ^2 values is due to sample size, as noted before these values will be more similar as sample size increases because the sampling distributions are asymptotically chi-squared. If one were to interpret these individual results, the number of course failures would best predict dropout/non-dropout status, followed by number of times retained in grade, number of suspensions, grade level, et cetera. Variable entry order, however, does affect results (see Appendix).

Table 4 indicates the main effects for the eight predictor variables and several criteria which are used to judge the significance of categorical independent variable entry in the logit model equation. A comparison of the hypothesized logit models with single predictors to the intercept model (independence model) is given by ΔC^2. A continuation of this table to include all 2-way interactions, 3-way interactions, 4-way interactions, et cetera would be required to determine the best set of predictor variables using the ΔC^2 criteria . Subsequently, one could compare the predictive efficacy of each logit model equation provided by the Δc value which indicates the percent above and beyond the *c* value for the intercept model. Calculation of the total number of logit model equations, i.e., 256, (2^m) is beyond the scope of this paper.

Table 3. Univariate L^2 and χ^2 on Dropout Status

Categorical Variable	L^2	χ^2	*df*	*p*
Gender	2.736	2.735	1	.09800
Ethnicity	52.859	52.481	3	.00001
Grade	99.508	97.137	5	.00001
Retained	139.872	133.628	1	.00001
Parent	34.084	32.189	1	.00010
Suspend	112.089	108.991	2	.00001
Economic	6.602	6.590	1	.01000
Course Failure	324.900	306.150	2	.00001

Note: L^2 and χ^2 are asymptotically chi-squared and become similar as sample size increases.

All Possible Subsets

The logit main effects and interaction effects model in study two would contain 256 equations in a saturated model. This is calculated by 2^m, where m=8 (Freund & Littell, 1991, *p.* 107). This does not take into account the fact that the order of entry for the categorical independent predictor variables would change the results. Many of the criteria for determining the best set of predictors have inherent problems. For example, the individual univariate Pearson chi-square or likelihood-ratio chi-square tests don't reflect interaction effects; the z-test of parameters in the logit model would change based on the order of entry in the equation and number of variables in the equation; the log-odds ratio because it is the exponentiation of the parameter estimate would also differ depending upon the order of entry and number of variables in the logit model equation; and the predictive efficacy (classification percent) is not necessarily a function of the significance of the parameters in the logit model. Consequently, the ΔC^2 (difference between -2logL values for null and hypothesized models) appears to be the most useful. A problem still remains in that SPSS and SAS do not provide a test for subsets of predictors nor do they generate all possible subset equations (Demaris, 1992, *p.* 68).

A new procedure, TABU (Drezner, Marcoulides, & Salhi, 1999), provides a solution to model selection in multiple regression which is directly applicable to logit modeling, and provides better results than a previously determined Mallows' *Cp* criteria (Schumacker, 1994). The Tabu program generates the *F*-ratio based on the L^2 and/or χ^2 value for all possible equations between the independence (null model) and the saturated model. Given a best model selection criteria of ΔC^2, one could easily pick the best set of categorical independent predictor variables. Subsequently, of the best ΔC^2 models, predictive efficacy could be compared (Δc).

Conclusions

In the first study, Table 2 indicated that a grade repeat status main effect was statistically significant accounting for 93% of the total modeled L^2. Locus of control main effect and interaction effects were not significant. [Please note that in Table 2, $L^2_{1\text{-}2}$, is the same as ΔC^2 in Table 4.] With only a few independent predictors one can easily hand calculate all of the possible subsets of equations. The entry order of independent predictor variables did have an impact on parameter estimates.

In the second study, Table 4 indicated that ethnicity, grade level, retained in grade, parent, suspensions, economic disadvantaged, and course failures main effects were statistically significant in the prediction of dropout/non-dropout status. Gender was not significant. A relative comparison of the ΔC^2 values for these main effects suggests that number of course failures followed by number of times retained in grade and number of suspensions would provide a possible best subset model. However, a researcher would not ultimately know the best subset model unless all possible subsets were calculated and compared on ΔC^2. The use of a Tabu search procedure to generate all possible subsets is therefore needed.

Educational Importance

The logit regression technique is **not** widely used in education even though it offers several advantages over the use of the chi-square statistic in analyzing categorical variables (Green, 1988). The type of variables used in these two studies are typical of the data recorded in school districts. A better understanding of this statistical technique, its applications, and interpretation will hopefully increase awareness of its value to educational researchers (Tabachnick & Fidell, 1989; Stevens, 1992).

Direct Correspondence to:
Randall E. Schumacker
College of Education
Matthews Hall 304
P.O. Box 311377
University of North Texas
Denton TX 76203-1337
E-Mail: rschumacker@unt.edu

Table 4. Logit Models: Main Effects Only

Main Effects	-2logL$_i$	ΔC^2	*df*	*p*	R^2_L	*c*	Δc
Intercept	2090.532					50%	
Gender	2087.796	2.736	1	.09800	.002	52%	2%
Ethnicity	2037.673	52.859	3	.00001	.034	59%	9%
Grade	1991.024	99.508	5	.00001	.064	61%	11%
Retained	1950.660	139.872	1	.00001	.089	62%	12%
Parent	2056.448	34.084	1	.00010	.022	54%	4%
Suspend	1978.443	112.089	2	.00001	.072	62%	12%
Economic	2083.930	6.602	1	.01000	.004	53%	3%
Course Failure	1765.632	324.900	2	.00001	.194	68%	18%

References

Anderson, C. (January, 1995). *Practical Applications for Logit Modeling*. Paper presented at the annual Southwest Educational Research Association, Dallas: Texas.

Demaris, A. (1992). *Logit Modeling Practical Applications*. Newbury Park: Sage Publications.

Drezner, Z., Marcoulides, G.A., & Salhi, S. (1999). Tabu search model selection in multiple regression analysis. *Communications in Statistics - Computation and Simulation*, 28(2).

Feinberg, S. E. *(1981). The analysis of cross-classified categorical data. (2nd Ed).*Cambridge, MA: MIT Press.

Freund, R.J. & Littell, R.C. (1991). *SAS System for Regression* (2nd Ed.). Cary, NC: SAS Institute, Inc.

Green, J. A. (1988). Loglinear Analysis of Cross-classified Ordinal Data: Applications in Developmental Research. *Child Development, 59*, 1-25.

Hinkle, D. E., McLaughlin, G. W. & Austin, J. T. (1988). "Using Log-Linear Model in

Higher Education Research" in *New Directions for Institutional Research*, Terenzini, P.T., (ed.) : San Francisco: Jossey-Bass.

Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*. NY: John Wiley.

Kennedy, J.J. (1992). *Analyzing Qualitative Data Log-Linear Analysis for Behavioral Research* (2nd Ed.). New York: Praeger Publishers.

Kleinbaum, D.G. (1992). *Logistic Regression: A self-learning text*. Springer-Verlag: New York.

Marasculio, L. A. & Busk, P. L. (1987). Loglinear Models: A way to study main effectsand interactions for multidimensional contingency tables with categorical data. *Journal of Counseling Psychology, 34*, 443-455.

Schumacker, R.E. (1994). A comparison of the Mallows *Cp* and principal component regression criteria for best model selection in multiple regression. *Multiple Regression Viewpoints, 21*(1), 12-22.

Stevens J. *(1992). Applied Multivariate Statistics for the Social Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tabachnick B. G. & Fidell, L. S. (1989). *Using Multivariate Statistics*. New York: Harper Collins Publishers.

APPENDIX
Variable Entry Order

FAILURE ENTERED FIRST

Null Model (Intercept Only): -2 Log Likelihood = 2090.5319

Hypothesized Model: -2 Log Likelihood = 1742.100

R^2: .206

Model Chi-Square (df=5): 348.432

Classification Overall: 71.15%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
FAILURES			128.0864	2	.0000	.2436	
FAILURES(1)	-2.0569	.1839	125.0690	1	.0000	-.2426	.1278
FAILURES(2)	-1.0140	.1612	39.5877	1	.0000	-.1341	.3628
RETAINED(1)	-.6551	.1681	15.1808	1	.0001	-.0794	.5194
SUSPEND			7.5456	2	.0230	.0412	
SUSPEND(1)	-.4718	.1730	7.4344	1	.0064	-.0510	.6239
SUSPEND(2)	-.2892	.2217	1.7012	1	.1921	.0000	.7489
Constant	1.9468	.1852	110.5476	1	.0000		

FAILURE ENTERED LAST

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
RETAINED(1)	-.6510	.1648	15.6055	1	.0001	-.0807	.5215
SUSPEND			7.5251	2	.0232	.0411	
SUSPEND(1)	-.4708	.1729	7.4136	1	.0065	-.0509	.6245
SUSPEND(2)	-.2887	.2218	1.6940	1	.1931	.0000	.7493
FAILURES	1.0301	.0910	128.0361	1	.0000	.2455	2.8012
Constant	-.1093	.2522	.1878	1	.6648		

Note: *df* = 4, Model Chi-square = 348.417