

Modeling Higher Education Processes with Multiple Logistic Regression: SAS Examples

Daniel Teodorescu, College of Charleston, Charleston, SC

ABSTRACT

This paper is intended to familiarize SAS users, research analysts and statisticians with the use of logistic regression in the study of higher education processes (i.e. college choice decisions, enrollment management, retention analysis, etc.).

LOGISTIC REGRESSION IN HIGHER EDUCATION

Many college outcomes and processes share the following two characteristics: 1) they are dichotomous in nature, and 2) they often are the result of an interaction between a series of student and institutional characteristics. Such outcomes/processes include, for instance, deciding to attend college, choosing a specific college over another, dropping out before completing a degree, graduating in a particular major and transferring from one institution to another.

By predicting/modeling these behaviors, institutional researchers can substantially affect the effectiveness of enrollment management and student retention programs at their institutions.

Logistic regression and only a few other statistical techniques (e.g. log-linear analysis, discriminant analysis, probit regression, and structural modeling) can conform to the specific nature of outcome measures such as enrollment, persistence, and degree attainment. Logistic regression is not new to the field of higher education research. Bishop (1977) used it to study college enrollment decisions and how responsive these decisions were to tuition and student aid packages. Manski and Wise (1983) employed logistic regression to examine the role of the Basic Educational Opportunity Grant (BEOG) in facilitating college choice and enrollment. They also used logistic regression to study the effects of SAT, family income, region, ethnicity and being a high school leader on the probability of applying for and being admitted to college. St John and St. John and Noell (1989) applied logistic regression to study the role of tuition on college attendance and to estimate the effect of student aid in facilitating college attendance on the part of minorities. Cabrera, Stampen and Hansen (1990) used this technique to explore the effects of ability to pay on persistence.

The example presented here draws from a retention study which attempts to predict the probability of leaving the institution for a cohort of entering students at an Engineering College.

Handling Input Data Sets

The unit of analysis in logistic regression can be either an individual case or a group of cases. In the later case, instead of one observation per case, the data set contains one observation per group, with a variable indicating the frequency for that group. A FREQ statement is used to provide the frequency variable. In our example, the data set has one observation per case. In SAS, columns are used to represent variables while rows stand for individual cases across variables (see Table 2). Table 2 shows the case number, the columns representing variable values for each independent variable (x1 through x11) and the observed probability of being retained (y1). The dichotomous dependent variable in our example is RETEN; students who were retained were coded as 1; those who left the

institution were coded as 0. By default SAS estimates the probability of leaving the institution (RETEN=0).

The set of independent variables in our example includes both dichotomous and continuous variables. The first group contains ethnicity (ETHNIC), sex (SEX), whether or not the student has changed majors (MAJCHA), whether or not the student has had a cooperative education semester (COOP), and whether or not the student lives in a campus dormitory (DORM). Continuous variables include first term GPA (FGPA), high school class rank (HSRANK), SAT verbal and math scores and the scores of two internally developed placement tests (TEST1 and TEST2). As to the coding scheme used, all dichotomous variables usually are coded as 0 and 1 to indicate the presence or absence of the classification condition. In the case of the variable COOP, the code 1 represent those students who had a semester of cooperative education experience while the code 0 represent those students who did not have such experience. When a categorical independent variable is made up of more than two categories, new dummy variables need to be created to represent the categories. For instance, if we were to recode ETHNIC (currently dichotomous with the levels Minority=1; Non-Minority=0) as a categorical variable with five levels (White, African American, Asian, Hispanic and Non-Resident) we would have to create four dummy variables: AFAM=1 for students who are African American and 0 otherwise, ASIAN=1 for Asian students and 0 otherwise, HISP=1 for Hispanic students and 0 otherwise, and WHITE=1 for White students and 0 otherwise.

Model Testing

In linear regression models, testing alternative models is usually done by examining whether or not the addition/deletion of variables significantly increases the proportion of variance explained in the dependent variable (or R Square). In logistic regression, there is a similar test based on the maximum likelihood function (coded "-2 LOG L" on the SAS printout and also called G Square or scaled deviance) associated with a particular model. However, unlike the R Square, the maximum likelihood function does not measure the proportion of variance explained, and therefore, has little value in deciding whether or not the model is a valid one. Alternative models are assessed by comparing the corresponding likelihood functions. In logistic regression, a better model yields a significantly smaller G Square. Table 1 shows the G-Square corresponding to our best fitting model.

Table 1. Criteria for Assessing Model Fit

Criterion	Intercept		Chi-Square For Covariates
	Only	Intercept and Covariates	
AIC	622.472	484.600	.
SC	626.638	534.585	.
-2 LOG L	620.472	460.600	159.872 (p=0.0001)
Score	.	.	122.579 (p=0.0001)

As with linear regression, building a model can follow a forward stepwise process or a backward stepwise process. Using the forward stepwise approach, individual variables or groups of variables are introduced sequentially, and the validity of the added variables is judged in terms of significant improvements of fit. In the backward stepwise approach, variables or groups of variables are deleted based on the extent to which their deletion significantly worsens the model. Our example illustrates the use of forward stepwise logistic regression:

```
PROC LOGISTIC DATA=COHORT;
MODEL RETEN=ETHNIC SEX/FORWARD;
MODEL RETEN= ETHNIC SEX SATMAT SATVER
HSRANK/FORWARD;
MODEL RETEN=ETHNIC SEX SATMAT SATVER
HSRANK FGPA TEST1 TEST2 MAJCHA
COOP DORM/FORWARD;
```

As shown in Table 4, three models (or steps) were sequentially estimated. The first model is intended to measure the extent to which demographic variables only impact one's probability to persist. The second model measures the impact of pre-college academic preparation variables, after controlling for demographic factors. The last model seeks to determine whether college experience variables affect the probability of persistence, after controlling for demographic and pre-college variables. Table 4 reveals that the addition of groups of variables increases the ability to predict retention, as implied by a reduction in the G-Square across the three models.

Assessing Goodness of Fit

Several methods are available to assess the how well a model fits the data. These include:

1. Pseudo "R Square"
2. Proportion of Cases predicted correctly
3. G Square/ df
4. Chi-Square for the overall fit

1. Pseudo "R Square". This indicator represents the proportion of error variance that an alternative model reduces in relation to a null model. The null model is usually specified as the one in which all the slopes with the exception of the intercept are set to zero. The formula for this indicator is as follows:

$$\text{Pseudo R-Square} = \text{Chi-Square} / (N + \text{Chi-Square})$$

where N is the total sample size. As is the case in linear regression with R Square, this indicator produces a conservative estimate of the reduction in error variance. For Model 3, in our example, the corresponding pseudo R Square shows that the model accounted for 22.8 percent in error variance ($\text{R Square} = 159.872 / (540 + 159.872)$).

2. Proportion of cases correctly predicted (PCP). This measure involves a comparison between the number of cases that the model predicted as being either 0 (i.e. not persisting) or 1 (i.e., persisting) against the total sample size. Using this measure, Table 3 indicates that the final model yielded correct predictions for 82 percent of the subjects. As the table shows, the PCPs increased as variables were added.

3. The G-SQUARE/df ratio. This indicator is similar to the chi-square test for LISREL models. The degrees of freedom are given by the following formula: $\text{df} = \text{sample size} - \# \text{ of parameters fitted}$. The degrees of freedom for our final model, for example, are 529 since we are estimating 11 parameters for a sample of 540 students. Since there are no established procedures to test the statistical significance of the G-SQUARE/df ratio, the rule of thumb is that a given model be accepted whenever the G-SQUARE/df ratio is less than 2.5. In our example, the final model meets such criterion ($\text{G-SQUARE/df} = .87$).

4. The chi-square statistics for overall fit. The measure the difference between the scaled deviance for the null model (in which only the intercept is fitted) and the scale deviance for the full model. A full model will be acceptable whenever there is a substantial reduction in chi-square. Unlike the measures of fit described so far, there are tests to assess whether the Chi-Square statistics for overall fit is statistically significant. This test is based on the fact that the difference in G-square numbers follows a Chi-Square distribution. For our final model, the observed Chi-Square = $620.472 - 460.600 = 159.872$ with corresponding $\text{df} = 11$ yielded a p-value less than .0001 which shows that the model fits the data.

Interpreting results

Table 5 presents the beta weights for our example. Like in linear regression, the sign associated to betas tell us the direction of the effect that a particular independent variable has on the dependent variable. For instance, the positive sign of the beta for the variable NCOOP indicates that students who had a cooperative education semester are more likely to persist, all other things being equal. However, unlike linear regression, the metric of the individual coefficients under logistic regression is in terms of logits rather than in terms of the original scale of measurement. For instance, in our example, Model 3 predicts that students who changed their major are 1.24 logit units less likely to leave the institution than are students who had the same major. Thus, the interpretation of the coefficients becomes rather impossible, particularly for categorical variables (where beta weights represent contrasts among categories expressed in terms of differences in logits).

To make possible the interpretation of the results, practitioners can use either one of the following two methods:

- 1) calculating/plotting predicted probabilities, and
- 2) computing delta-p statistics.

The latter is accomplished by measures of overall change in the dependent variable.

Calculating/plotting predicted probabilities. As in linear regression, predicted probabilities for each case (or subject) can be computed by multiplying the values of each variable by its corresponding Beta weights and adding the resulting products. The logistic equation in Table 3 can be summarized as follows:

$$\begin{aligned} \text{Logit } i_{th} = & p/(1-p) = 2.2483 + 0.1956\text{RACE} + .1783\text{SEX} \\ & -.0029\text{SATMAT} + .0010\text{SATVER} + .0057\text{HSRANK} \\ & -.7925\text{FGPA} - .0180\text{TEST1} + 0.0079\text{TEST2} \\ & -1.2442\text{MAJCHA} -2.5852\text{COOP} + .3871\text{DORM} \end{aligned}$$

Where Logit i_{th} represents the predicted natural logarithm value or the logit for the i_{th} subject. The predicted value for the first student in our example (case 1 in Table 2), for instance is:

Logit1=-4.2865

By using the formula for logit transformation, one can calculate the corresponding predicted probability of leaving the institution:

$$P1 = \exp(-4.2865) / (1 + \exp(-4.2865)) = .014$$

In SAS, PROC LOGISTIC calculates predicted probabilities and logits and these results may be saved in an output data set, from which plots can be made.

```
PROC LOGISTIC DATA= COHORT;
  MODEL RETEN=ETHNIC SEX SATMAT SATVER
  HSRANK FGPA TEST1 TEST2 MAJCHA COOP
  DORM;
  OUTPUT OUT=RESULTS P=PREDICT L=LOWER
  U =UPPER XBETA=LOGIT;
  PROC PRINT DATA=RESULTS;
```

The OUTPUT statement produces a data set containing logit values for each subject, corresponding predicted probabilities and confidence limits (UPPER, LOWER) for those probabilities. To plot the predicted probabilities from the RESULTS data set, we select all observations with RETEN=1. A plot can be done as a bar chart with PROC GCHART or as a line graph with PROC GPLOT.

Aside from predicting whether a given subject will engage in a particular behavior, practitioners can also estimate probabilities attributable to a particular variable that may be of interest to them. For instance, one might be interested in assessing the effects of changing majors across gender and ethnicity.

```
PROC SORT DATA=RESULT;
  BY SEX MAJCHA SEX ETHNIC;
```

```
PPROC GPLOT;
  PLOT PREDICT * ETHNIC=MAJCHA/ NOLEGEND;
  BY SEX;
```

Delta-p statistics. These indicators measure changes in the probability resulting from a unit change in the independent variable and are used to estimate the overall change in the dependent variable. They can also be employed for assessing the relative size of these changes across variables, provided that the independent variables have the same metric. Petersen's (1985) formula for Delta-p is:

$$\text{Delta-p} = \exp(L1) / [1 - \exp(L1)] - P0$$

Where P0 is the sample mean of the dependent variable. In our example, the probability of persistence for the whole sample is .680 (P0=.680)

L1=L0 + B represents the logit after the unit change in the variable under consideration. L0 is the natural logarithm of $[P0/(1-P0)]$.

The incremental effect attributable to MAJCHA can be calculated as follows:

$$L0 = \ln[P0/(1-P0)] = \ln[.680/(1-.680)] = \ln(2.125) = .7538$$

$$L1 = L0 + B_{majcha} = .7538 - 1.2442 = -.4904$$

Therefore the logit before the change is L0= .7538, and after the unit change is L1=-.4904. The Delta-p for MAJCHA yields the probability value of .63.

$$\text{Delta-p} = \exp(.7538) / [1 + \exp(-.4904)] - .680 = 1.3179 - .680 = .6379$$

The model predicts that changing majors decreases the probability of leaving the institution by 63.8 percentage points.

In the case of continuous independent variables, delta-p represents the incremental effect on the outcome variable resulting from a unit change in the independent variable.

REFERENCES

- Bishop, J. (1977). The effects of public policies on the demand for higher education. *Journal of Human Resources*, 12, 285-307.
- Cabrera, A.F., Stampen, J.O. and Hansen, W.L. (1990). Exploring the effects of ability to pay on persistence in college. *Review of Higher Education*, 13(3), 306-336.
- Manski, C.F. and Wise, D.A. (1983). *College Choice in America*. Cambridge, MA: Harvard University Press.
- Petersen, T (1985). A comment on presenting results from logit and probit models. *American Sociological Review*, 50(1), 130-131.
- St. John, E. P. (1991). What really influences minority attendance? Sequential analyses of the High School and Beyond Sophomore Cohort. *Research in Higher Education*, 32(2), 141-158.
- St. John, E.P. and Noell, J. (1989). The effects of student aid on access to higher education: An analysis of progress with special consideration of minority enrollment. *Research in Higher Education*, 30(6), 563-581.

Table 2. Layout of Data for Logistic Regression

Case#	ETHNIC (x1)	SEX (x2)	SATMAT (x3)	SATVER (x4)	HSRANK (x5)	FGPA (x6)	TEST1 (x7)	TEST2 (x8)	MAJCHA (x9)	COOP (x10)	DORM (x11)	RETEN (y)
1	1	1	760	650	4	3.256	60	70	1	0	1	1
2	0	0	640	620	8	3.230	72	84	0	1	1	1
3	0	1	660	550	20	2.456	56	62	1	1	0	0
4	1	1	560	540	18	2.810	57	66	0	1	0	1
.	.	.	.	.	.	.	.	.	.	.	.	.
540	0	1	530	620	15	3.110	64	57	0	0	1	0

Table 3. Effects of Demographic, Pre-College and College Variables on the Probability of Leaving the Institution.

Variable (Beta)	Step1 (Beta)	Step2 (Beta)	Step 3 (Beta)
ETHNIC	.3476	.5622	0.1956
SEX	.3184	.3424	0.1783
SATMAT		-0.00401	-0.00294*
SATVER		0.00236	-0.000981
HSRANK		0.00247	0.00574
FGPA			-0.7925***
TEST1			-0.0180
TEST2			0.00797
MAJCHA			-1.2442***
COOP			-2.5852**
DORM			0.2871
Intercept	-1.2188	-0.4202	2.2483
G Square	672.064	638.134	460.600
Df	538	535	529
G Square/df	1.249	1.193	0.871
Pseudo R Sq	.009	.037	.228
PCP	36.5	62.5	82.4
Chi Square	5.256	20.703***	159.872***

*p < .05 ** p< .01 *** p<.001

Table 4. Effects of Adding Variables on the Fit of the Model.

Model	df	G Square	Change in df	Change in G Square	Improvement Of fit p-value
1. Demographic variables	538	672.064			
2. Adding pre-college variables	535	638.134	df1-df2=3	33.930	.0009
3. Adding college variables	529	460.600	df1-df2=6	177.534	.0001

Table 5. Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCEPT	2.2483	0.6081	13.6717	0.0002	.
ETHNIC	0.1956	0.2691	0.5280	0.4674	0.052764
SEX	0.1783	0.3307	0.2906	0.5898	0.036976
SATMAT	-0.00294	0.00143	4.2543	0.0392	-0.252118
SATVER	0.000981	0.00163	0.3620	0.5474	0.075543
HSRANK	0.00574	0.00412	1.9354	0.1642	0.102089
FGPA	-0.7925	0.1479	28.6997	0.0001	-0.386104
TEST1	-0.0180	0.0221	0.6661	0.4144	-0.078867
TEST2	0.00797	0.0101	0.6246	0.4293	0.073806
MAJCHA	-1.2442	0.2353	27.9492	0.0001	-0.488304
COOP	-2.5852	0.7152	13.0657	0.0003	-0.761227
DORM	0.3871	0.2412	2.5774	0.1084	0.101089

Association of Predicted Probabilities and Observed Responses

Concordant	=82.4%	Somers' D	= 0.650
Discordant	=17.4%	Gamma	= 0.652
Tied	= 0.2%	Tau-a	= 0.299
(52020 pairs)		C	= 0.825