USING LOGISTIC REGRESSION TO PREDICT CUSTOMER RETENTION

Andrew H. Karp
Sierra Information Services, Inc.
San Francisco, California USA

Logistic regression is an increasingly popular statistical technique used to model the probability of discrete (i.e., binary or multinomial) outcomes. When properly applied, logistic regression analyses yield very powerful insights in to what attributes (i.e., variables) are more or less likely to predict event outcome in a population of interest. These models also show the extent to which changes in the values of the attributes may increase or decrease the predicted probability of event outcome.

This approach to "pattern recognition" or "data mining" is particularly well suited to applied statistical analyses of consumer behavior. Logistic regression models are frequently employed to assess the chance that a customer will: a) re-purchase a product, b) remain a customer, or c) respond to a direct mail or other marketing stimulus. This paper discusses how logistic regression, and its implementation in the SAS/STAT™ module of the SAS® System, can be employed in these situations. We will focus on concepts and implementation, rather than the statistical theory underlying logistic regression; the reader is encourage to consult the sources listed at the end of the paper for more information about the statistical theory and other details of logistic regression modeling.

What is Logistic Regression?

Logistic regression refers to statistical models where the dependent, or outcome variable, is categorical, rather than continuous. The *logistic function* "maps" or "translates" changes in the values of the continuous or dichotomous independent variables on the right-hand side of the equation to increasing or decreasing probability of the event modeled by the dependent, or left-hand-side, variable. These statements highlight the key difference between logistic regression and "regular" (more formally, *ordinary least squares*) regression: in logistic regression, the predicted value of the dependent variable being generated by operations on the right-hand-side variables is a *probability*. But, in ordinary least squares regression we are predicting the population mean v*alue* of the dependent variable at given levels (i.e., values) of the independent variable(s) in the model.

Economists frequently call logistic regression a "qualitative choice" model, and for obvious reasons: a logistic regression model helps us assess probability which "qualities" or "outcomes" will be chosen (selected) by the population under analysis. When proper care is taken to create an appropriate dependent variable, logistic regression is often a superior (both substantively and statistically) alternative to other tools available to model event outcomes. The SAS System's implementation of logistic regression techniques include a wealth of tools for the analyst to use to first construct a model, and then test its ability to perform well in the population from which the data under analysis are assumed to be a random sample.

Before proceeding further a clarification is warranted as to the nature of the dependent variable. In many analytic scenarios the customer (or other "unit of analysis") may have more than one choice available to them. For example, during "open enrollment" a health plan participant may be offered a chance to: maintain their current health care plan, choose one with a richer benefits mix, select one with fewer benefits, or select a plan from a competing provider. In this situation a *multinomial logistic regression model* might be employed, rather than a *binary* or *dichotomous logistic regression model.* While multinomial models can be implemented in both PROC LOGISTIC and PROC GENMOD, they are beyond the scope of this discussion, which will be limited to models where a binary (i.e., "one outcome or the other") outcome is to be analyzed.

Why is Logistic Regression Often a Superior Alternative to Other Methods to Analyze Event Outcome?

Other approaches besides logistic regression are often proposed to develop models of event outcome or qualitative choice. These include: using ordinary least squares regression with a binary (zero/one) dependent variable, discriminant analysis, and probit analysis. While each of these competing methods are applicable to the discrete choice modeling scenario, the logistic regression model is most often a superior approach for a number of statistical (as well as substantive reasons), including:

- The logistic regression equation limits generation of the predicted values of the dependent variable to lie in the interval between zero and one; whereas OLS regression often results in values of the dependent variable take on values of less than zero or greater than one, which are substantively irrelevant and have no "interpretative" value.

- A simple transformation (exponentiation) of the logistic regression model's parameters leads to an easily interpretable and explainable quantity: the odds ratio. In more recent releases of SAS/STAT™ software the odds ratio is printed for each independent variable (parameter) in the model.

- A number of useful tests for assessing model adequacy and fit are available for logistic regression models. These include: measures similar to the "coefficient of determination" in OLS regression; a generalized test (Hosmer-Lemeshow) for lack of model fit, and the ability to develop tables which show the proportion of cases under analysis which have been correctly classified; false positive and negative rates; and the sensitivity and specificity of the model. Tests are also available for "influential" and other ill-fitting observations.

- Parameter estimates generated from a logistic regression model can be applied in a simple Data Step to the "population of interest," this "scoring," or creating a probability of event outcome for each "member" of the population. This "score" can then be used to select subsets of the population for various "treatments" as may be appropriate to the substantive issue under analysis.

Implementation of logistic regression models in the SAS System in PROC LOGISTIC is very similar to how OLS regression models are implemented in PROCs REG and GLM. If you are already familiar with how to perform OLS regression in PROC REG then learning how to use PROC LOGISTIC for binary outcome modeling is a straightforward task. As with PROCs REG and GLM, separate output data sets containing predicted values and parameter estimates can be created for subsequent analysis or "scoring."

Construction of the Dependent Variable: a Critical Consideration

In many analytic situations where logistic regression is the method of choice the analyst has several—or more—independent variables to use in the modeling process. These variables may be the result of some "naturally occurring" or "observable" aspect of the behavior of the units under analysis, or be constructed from other variables. For example, an analyst developing a model predicting re-enrollment in a health insurance plan may have data for each member's interaction with both the health plans administrative apparatus and health care utilization in the prior "plan year." The analyst can, with appropriate prior knowledge of the "substantive" relevance of the data at hand, as well as knowledge of how to use the SAS System's Data Step to operate on variables in a SAS data set, obtain or construct variables which can be employed in the modeling process. These variables may include: number of times member called the health plan for information, number of physician office visits, whether or not the member changed primary care physicians during the previous "plan year," and answers to a customer satisfaction survey.

Using this example, it is clear that the analyst has an array of attributes (variables) from which to choose in developing models. What is not yet available is the variable which can employed as the "outcome" or "dependent" variable. In logistic regression analyses it is often the analyst's responsibility to *construct* the dependent variable based on an agreed-upon definition of what constitutes the "event of interest" which is being modeled.

From a statistical standpoint, the dichotomous, nominal level dependent variable must be both "discriminating" and "exhaustive." The variable has two levels (i.e., it is dichotomous), and the values of the variable can choose (or "discriminate") between observations in the population of interest. Finally, the dependent variable is considered "exhaustive" because each individual observation can be assigned to one, and only one, of its values.

In most consumer behavior studies the dependent variable represents the analyst's "operationalization of the construct" of interest, rather than a phenomenon which naturally occurs in the "environment" from which the values of independent variables measured. Creation of the dependent variable therefore depends on the analyst's own understanding of the consumer phenomenon of interest, as well as any internal "rules" their organization follows when defining the outcome of interest. Again returning to the health care re-enrollment example from above, a health plan's management team may define "attrition" or "failure to re-enroll" as situations where a member fails to return the re-enrollment card within 30 days of its due date. Or, in a response modeling scenario, a direct mail firm may define "non-response" to an advertisement as failure to respond within 45 days of mailout.

These "rules" are carefully implemented by the analyst during construction of the dependent variable in a Data Step. This often requires clear communication between the analyst and others in their organization as to how the "rules" will be translated in to SAS programming language statements that will construct the dependent variable. In many applied situations construction of the dependent variable, once the rules for it have been agreed-upon, requires extensive Data Step manipulation of variables and data sets. Failure to respond to a mail circular may, for example, be determined only by match-merging two data sets, one with containing customer information for all persons to whom the mailing was sent, and another with just those who responded to it within the agreed-upon time frame. If a customer is found in the "mailed to" data set, but not found in the "responses" data set, the analyst may then use appropriate Data Step coding to classify the customer as a "non respondent." Similarly, a customer who has been "found" on both data sets will be classified as a "respondent" to the mail campaign. From an applied standpoint, appropriate construction of the dependent variable cannot be overemphasized. Failure on the analyst's part to correctly understand both the "construct" (i.e., event) whose probability is to be modeled, as well as how to "code the rules" which "operationalize" the construct may result in models with limited substantive as well as statistical usefulness.

Implementing and Interpreting a Logistic Regression Model

We now turn to the implementation and interpretation of a logistic regression model. PROC LOGISTIC, in the SAS/STAT™ module, contains the tools necessary to apply a logistic regression model to a data set and assess its results. As with its ordinary least squares counterpart, PROC REG, the MODEL Statement in PROC LOGISTIC contains the dependent variable to the left of an equals sign (=) and the name(s) of the independent variable(s) on the right hand side of the equals sign. Options are placed to the right of a slash or stroke mark (/) following the last independent variable in the model. As with PROC REG, the optional OUTPUT statement in PROC LOGISTIC can be used to create an output SAS data set. This topic will be explored in detail below.

Output from successful execution of a PROC LOGISTIC 'step' in a SAS Software program yields a wealth of information about both the overall fit of the model, as well as the 'statistical significance' of each of its parameters. The –2LOGL test is analogous to the "global F test" in OLS regression, and is used to determine the overall impact of the independent variables on predicting increasing or decreasing probability of event outcome. The 'local' chi-square tests are the counterparts to the 'local' t-tests in OLS regression, as the test whether or not the associated parameter is 'statistically significant.'

As mentioned previously, exponentiation of the parameter estimates yielded by PROC LOGISTIC yields a very useful and informative value: the odds ratio. This quantity represents the increasing or decreasing probability of event outcome per unit change in the value of the independent variable. The odds ratio is therefore a very valuable substantive, as well as statistical, outcome of applying a logistic regression model to a customer retention or other qualitative choice scenario. The odds ratio is: a) easily calculated (and printed by default by PROC LOGISTIC in recent releases of SAS/STAT™ software), and, b) easily interpreted by non-quantitative personnel who might use the results of a logistic regression model to plan a marketing or other campaign Since customer retention efforts also have an economic component (e.g., "how much will it cost to keep a client?"), articulation of the results of a model in the context of the parameter's odds ratios makes it much easier for all decision makers in the customer retention effort to grasp the substantive impact of the parameters on the phenomenon under study.

In my own experience, telling a financial institution's marketing manager "if we add a credit card to a household's product portfolio, in addition to their checking and saving accounts, the probability we will keep them as a customer for another year increases by x percent," or being able to explain to an HMO's sales team that "for every additional year the member has been with our HMO, the chance they will re-enroll increases by y percent" permits the substantive relevance of a logistic regression model to be readily understood by those who will use its results to plan or modify strategy.

Customized odds ratios, as well as confidence intervals for odds ratios are available from PROC LOGISTIC using the UNITS and RISKLIMITS options. By default, invocation of the RISKLIMITS option generates 95% confidence intervals around the odds ratios; users can specify other intervals using the ALPHA option in conjunction with the RISKLIMITS option.

Selecting the "Optimal Subset" of Independent Variables

In many modeling situations the analyst may have many variables to consider for inclusion in their model(s). In addition to relying on the "substantive significance" of the variables from which to choose, PROC LOGISTIC permits implementation of automated variable subset selection tools such as forward, backward and stepwise generation of "optimal subsets." As with any other type of modeling effort, the results selected as "best" via an automated subset selection process implemented in PROC LOGISTIC should be carefully examined in both the statistical and "substantive" relevance of the variables chosen. In addition, both multicollinearity and 'influential observations' are issues in logistic regression just as they are with ordinary least squares regression.

Various measures for both conditions are available by specifying the INFLUENCE and IPLOTS options in PROC LOGISTIC. The reader is directed to pp. 467-470 of SAS Institute publication “SAS/STAT Software: Changes and Enhancements Through Release 6.12” (full citation below) for more information on diagnostic tests implemented via specification of these options.

Choosing from Competing Models

As with other forms of statistical modeling, the logistic regression modeling effort frequently yields more than one “competing” model of the phenomenon under study. Fortunately, PROC LOGISTIC provides several tests which help assess: a) how well the independent variables “fit” the model; and, b) the model’s ability to correctly predict the outcome modeled by the dependent variable.

Among the measures of model fit available in PROC LOGISTIC are the Hosmer-Lemeshow goodness of fit test (generated by the LACKFIT option in the MODEL statement) and several measures similar to the “coefficient of determination” obtained in ordinary least squares regression. These “r-square like” statistics are generated when the RSQUARE option is invoked. These measures generalize the “coefficient of determination” to the categorical modeling situation and help assess how well the independent variables “fit” the model.

Another approach is to assess the sensitivity (proportion of observations with the condition of interest which are predicted by the model to have the condition of interest) and/or specificity (proportion of cases without the condition of interest which are predicted by the model to NOT have the condition of interest). In addition, the false positive and false negative rates of several competing models may also be of interest. These and other common measures of model adequacy are portrayed in “classification tables” generated by PROC LOGISTIC when the CTABLE option is specified. If the analyst has a specific (or range) of prior probabilities of event occurrence for which they want a classification table generated, the PPROB option results in limiting generation of the classification table to just those desired by the analyst.

Implementing the Results of a Logistic Regression Model

Once a model predicting customer retention has been selected (usually on the basis of a combination of statistical and substantive considerations), attention usually turns to the implementation of the model in the organization. In my experience, there are often two distinct aspects to how an organization applies the results of a model.

The first aspect is dissemination of the model’s results to others in the organization. Employees may be told, for example, about the results of the model (in a non-statistical fashion) via internal communication channels. Managers may be tasked to implement the results of the model via changes in strategy or on going programs. For example, the credit card marketing department of a financial institution may decide to increase its offers of credit cards to (creditworthy) customers if the results of a model suggest that adding this product to a customer’s “portfolio” increases the chance of retention. In one organization where I have consulted, management compensation plans were changed to provide financial incentives to those line managers who were successful in maintaining or improving customer retention rates. As part of the compensation plan’s implementation managers attended training sessions where the retention model was discussed in the context of the “drivers of customer retention” and how these managers were responsible for employee performance on these “drivers.”

A second, and very common, aspect to model implementation is to “score” or apply the selected model’s parameters to every customer of the organization commissioning the model. Most customer retention modeling efforts take place with a (hopefully) random sample of a larger population of interest. Once a model has been chosen, its parameters are applied to the entire population of interest, thereby obtaining a “probability to stay (or leave)” score for each customer. Once generated, these scores can then be used to identify groups of customers who are more (or less) likely to stay (or leave), and marketing efforts tailored accordingly. The OUTEST option in the PROC LOGISTIC statement will generate an output SAS data set containing model parameters. The variables in this data set can be used in subsequent Data Steps to implement the scoring process.

Summary and Conclusions

Logistic regression is a very powerful tool in building models of customer retention. SAS System software implements this tool in both PROC LOGISTIC in the SAS/STAT™ module and the forthcoming Enterprise Miner™ product. When applied properly, logistic regression models can yield powerful insights in to why some customers leave and others stay. These insights can then be employed to modify organizational strategies and/or assess the impact of the implementation of these strategies.

As with any other form of statistical modeling, care must be taken to carefully construct variables representing the phenomenon of interest and to select candidate independent variables that are “substantively relevant” to the issues under analysis.

Finally, the dynamic nature of most “systems” of customer choice suggests that periodic re-assessment of the selected model is highly appropriate. As customer demands and needs, as well as the environment" in which the organization competes, change, so must the models the organization employs to predict customer retention.

Note: SAS, SAS/STAT and Enterprise Miner are trademarks of SAS Institute, Cary, NC. USA.

References

Berry, Michael J.A., and Gordon Linoff, *Data Mining Techniques For Marketing, Sales and Customer Support,* Wiley, 1997

Groth. Robert, *Data Mining: a Hands-On Approach for Business Professionals,* Prentice-Hall, 1998

Hosmer and Lemeshow , *Applied Logistic Regression,* Wiley:, 1989

Kennedy, Ruby L, et,al, *Solving Data Mining Problems Through Pattern Recognition,* Prentice-Hall, 1997

The author can be contacted at:

Sierra Information Services, Inc.
1489 Webster Street
Suite 1308
San Francisco, California 94115
415/441-0702
SierraInfo @ AOL.COM
www.SierraInformation.com

SAS Institute, Inc: *SAS/STAT Software, Volume 2: the LOGISTIC Procedure*

SAS Institute, Inc.: *SAS/STAT Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07*

SAS Institute, Inc.: *SAS/STAT Software: Changes and Enhancements Release 6.10*

SAS Institute, Inc.: *Logistic Regression Examples Using the SAS System* (1995)

SAS Institute, Inc: *SAS/STAT Software: Changes and Enhancements through Release 6.12* (1997)

Weiss, Sholom M., and Nitin Indurkhya, *Predictive Data Mining: A Practical Guide,* Morgan Kaufmann, 1998