Association Metrics

Calculating the Probabilities of Member Engagement

by

Larry J. Seibert, Ph.D.

Binary logistic regression is a regression technique that is used to calculate the probability of an outcome when there are only two possible options, such as whether or not a member will renew his/her membership. There are a number of behaviors that drive member engagement for which this type of research is applicable, such as whether or not the individual member will:

1. renew his/her membership
2. attend conventions and conferences
3. attain/maintain a professional certification
4. attend continuing education classes
5. participate in webinars
6. attend local functions
7. recruit new members
8. join special interest groups/committees
9. use a particular member benefit
10. visit the website
11. interact with social media
12. donate to a foundation
13. contact the call center for assistance
14. volunteer at the national or local level
15. experience a problem with the organization
16. transition from a student membership to a professional membership.

The purpose of this paper is show how binary logistic regression can be used to develop predictive models for each of the desirable outcomes for the organization, and in the process, identify the predictors of those outcomes. This research can be used to identify those individuals with high probabilities of participating in an organization's programs and aid in the marketing of those programs to members.

In addition, the model can also determine those individuals with low probabilities of desirable outcomes, such as renewing his or her membership, or not transitioning from a student membership to a professional membership. This will identify individuals for whom the organization can preemptively target with retention efforts.

(While there are several software packages that can generate binary logistic regression models, the examples presented in this paper are generated from IBM SPSS. For

illustrative purposes, the example presented throughout this paper will be whether or not members will attend the organization's annual convention.)

How it Works

Binary logistic regression is similar to other regression techniques in that it uses independent variables (predictors) to predict the outcome of a dependent variable (e.g. convention attendance). Where it differs from other types of regression, such as multiple regression, is that in binary logistic regression, the dependent variable has only two possible outcomes (e.g. attend/not attend), and what is being predicted is the probability, or the odds, that an individual will be placed into one group or the other. Probabilities can range from 0 to 1.

Dependent Variable

In the binary logistic regression procedure, the dependent variable is the outcome to be predicted (e.g. annual convention attendance). For every member who exhibits the outcome (attended the annual convention), the value of "1" is assigned, and for every member who does not exhibit the outcome, the value of "0" is assigned.

The binary logistic regression procedure will assess how well the set of predictor variables does in predicting the dependent variable. The function will generate an odds ratio for each independent variable, which represents the change in odds of being in one of the categories of outcome when the value of a predictor increases by one unit.

Additionally, the probability of attending the convention, based on the set of independent/predictor variables, will be generated for each individual.

Independent Variables

Independent variables can be any information that is available to the organization about its members that the researcher would like to include in the model. Any information that is stored internally by the organization (e.g. the number of years an individual or organization has been a member, continuing education classes attended, volunteer hours worked) or information that can be obtained from a survey (e.g. likely to recommend the organization, likely to renew their membership, perceived value of their membership, reasons they joined/belong, who/what influenced them to join, satisfaction with the direction of the organization, etc.), and demographic information (age, sex, race, geographic region, formal education, job title) are all variables that can be used as predictor variables. For binary logistic regression, independent/predictor variables can be continuous, ordinal, nominal, or a mix of all three.

Ideally, independent/predictor variables should be strongly related to the dependent variable, but not strongly related to the other independent variables. One of the best ways to explore which variables are likely candidates for a binary logistic regression model is to run a correlation analysis between the dependent variable and potential independent variables.

Correlations

		Convention Attendance	Years a Member	Years in the Profession	Local Volunteer	Percent of Income
Convention Attendance	Pearson Correlation	1	.417**	.399**	-.517**	.038
	Sig. (2-tailed)		.000	.000	.000	.109
	N	1832	1832	1823	1832	1810
Years a Member	Pearson Correlation	.417**	1	.798**	-.295**	.090**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	1832	1832	1823	1832	1810
Years in the Profession	Pearson Correlation	.399**	.798**	1	-.270**	.121**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	1823	1823	1823	1823	1802
Local Volunteer	Pearson Correlation	-.517**	-.295**	-.270**	1	-.065**
	Sig. (2-tailed)	.000	.000	.000		.006
	N	1832	1832	1823	1832	1810
Percent of Income	Pearson Correlation	.038	.090**	.121**	-.065**	1
	Sig. (2-tailed)	.109	.000	.000	.006	
	N	1810	1810	1802	1810	1810

**. Correlation is significant at the 0.01 level (2-tailed).

In the correlation table above, Years a Member and Years in the Profession, are both candidates for inclusion in the binary logistic regression model, as both have “relatively” high correlation coefficients with Convention Attendance (.417 and .399 respectively). Not surprisingly, they are strongly correlated to each other (coefficient of .798), since a longer time as a member of this individual professional organization is expected to correlate with a longer time spent working in the profession.

Since Years a Member has a larger correlation coefficient with Convention Attendance (.417) compared to the coefficient between Years in the Profession and Convention Attendance (.399), Years a Member will be used in the regression model.

The dependent variable (Convention Attendance) is coded 0, 1, with 0 for not attending and 1 for attending. The positive sign for the Years a Member coefficient means that a higher probability of attending the convention is associated with individuals who have been members longer.

The correlation coefficient for Local Volunteer (-.517) has a high absolute value, and will also be included in the regression model. The negative sign for this coefficient indicates an inverse relationship between these two variables. Because the dependent variable (Convention Attendance) is coded 0, 1, with 0 for not attending and 1 for attending, there is a higher probability of attending the convention associated with smaller values of the variable Local Volunteer. Since the Local Volunteer variable is coded as 1 = current volunteer, 2 = former volunteer and 3 = never volunteered, the negative sign for this coefficient seems reasonable. It is plausible that members who are current volunteers have a higher probably of attending the annual convention than members who have never been involved as a volunteer.

The variable, Percent of Income, which measures the percent of the individual's household income that is derived from working in this profession, has a very small correlation coefficient (.038) and is not statistically significant (p=.109). Therefore, Percent of Income will not be included as a predictor variable in the model.

All information that can theoretically explain whether a member is likely or unlikely to engage in a particular behavior should be tested to ensure that no predictor variables are excluded from the model. Once the correlation analysis has generated a short list of predictor variables, the next step is to run the binary logistic regression procedure with the dependent variable and each independent/predictor variable alone in order to understand how each independent variable interacts with the dependent variable.

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1832	100.0
	Missing Cases	0	.0
	Total	1832	100.0
Unselected Cases		0	.0
Total		1832	100.0

a. If weight is in effect, see classification table for the total number of cases.

The Case Processing Summary on the previous page shows the number of cases included in the procedure. When there are multiple independent variables in the model, only observations that contain a value for every independent variable in the model will be included in the analysis.

The Dependent Variable Encoding below confirms that individuals who did not attend the convention are coded as 0 and those who attended are coded as 1.

Dependent Variable Encoding

Original Value	Internal Value
Did not attend a Convention	0
Attended a Convention	1

Block 0 is the model absent any predictor variables. The Observed numbers indicate those who actually attended or did not attend, while the Predicted numbers are those predicted by the model.

Based on the observed numbers in this example 1284 individuals did not attend the convention (70.1%) and 548 members did attend (29.9%). Because there are no predictor variables, the model will predict that no one will attend (see numbers in the Did not attend a Convention column) because that produces a higher percentage of correct classifications than if it were to predict that everyone would attend. At this stage, there is no information available from which to make predictions.

Independent variable(s) will be added to increase the predictive ability of the model that has no independent variables.

Block 0: Beginning Block

		Predicted		
		Convention Attendance		
Observed		Did not attend a Convention	Attended a Convention	Percentage Correct
Convention Attendance	Did not attend a Convention	1284	0	100.0
	Attended a Convention	548	0	.0
Overall Percentage				70.1

Block 1 indicates the inclusion of the independent variable. The Omnibus Tests of Model Coefficients show that the model that includes the independent variable is significant (Sig. = .000)

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	315.054	1	.000
	Block	315.054	1	.000
	Model	315.054	1	.000

Binary logistic regression does not have an R^2 similar to multiple regression. What it has is what statisticians refer to as "pseudo" R^2. The two examples below are the Cox & Snell R^2 and the Nagelkerke R^2. While both statistics are typically shown in reports, the Nagelkerke R^2 is preferred as it is scaled to have an upper limit of 1, while the scale on the Cox and Snell R^2 has an upper limit of .75, which generates a lower number.

The Nagelkerke R^2 of .224 for a one independent variable model is very good. If this were an R^2 for a multiple regression model, we could say that this independent variable explains 22.4% of the variation of the dependent variable.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1920.435[a]	.158	.224

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

		Predicted Convention Attendance		
Observed		Did not attend a Convention	Attended a Convention	Percentage Correct
Convention Attendance	Did not attend a Convention	1164	120	90.7
	Attended a Convention	357	191	34.9
Overall Percentage				74.0

The classification table on the previous page shows the predictions that are made with the one independent variable model. The predictive accuracy of the model has increased from 70.1% with no independent variables in the model, to 74.0% with the Years a Member variable added to the model. The numbers that are circled in the classification table indicate the number of observations that are correctly classified in this model. The numbers not circled represent the number of observations that are incorrectly classified.

"The sensitivity of the model is the percentage of those who attended the convention that were correctly classified (34.9%). The specificity of the model is the percentage of those who did not attend the convention and were correctly classified (91.7%).

The positive predictive value is the percentage of cases that the model classifies as having attended the convention, and who actually attended

(1) 191 / (191 + 120) = 61.4%.

The negative predictive value is the percentage of individuals predicted by the model to have not attended the convention, and, in fact, did not attend the convention

(2) 1164 / (1164 + 357) = 76.5%" (Pallant, J., 2016).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Years a Member	.116	.007	262.376	1	.000	1.123
	Constant	-2.030	.095	451.730	1	.000	.131

a. Variable(s) entered on step 1: Years a Member.

The first item to check in the table above is whether or not the independent variable is significant. The Sig. of .000 shows that this independent variable is significant. The B value is analogous to the B value in multiple regression – it represents the incremental change as the value of the independent variable changes by one unit. However, where the B in multiple regression is used to calculate the predicted value of the dependent variable, the B in binary logistic regression is used to calculate the change in the logit, which can then be used to calculate the probability of being assigned to the "attend the convention group".

The example below shows how to use the B value of the independent variable and the constant to calculate the probability of a person attending the annual convention, who has been a member for ten years.

(3) .116 (10) + (-2.030) = - .87

(4) $e^{-.87}$ = .41895

(5) .41895 / (1 + .41895) = 29.5%

By substituting the number of years a person has been a member in equation 3 above, the probability of attending the annual convention for each year of membership can be calculated. For example, for individuals who have been a member for 20 years, their probability of attending would be 57%.

An assumption that is made is that the incremental change in B (.116) is constant throughout the full range of values for Years a Member. However, that may not be true, and there is a simple way to test for this. This can be done by collapsing the number of years into categories and treating the categorical variable as a nominal variable – one in which the values of the variable serve to differentiate one group from another and do not assume a relationship among groups.

The new variable, Years a Member Category, will contain the following values:

1 = 0- 2 years
2 = 3-5 years
3 = 6-10 years
4 = 11-15 years
5 = 16-20 years
6 = 21-30 years
7 = 31 + years.

Any continuous variable can be treated as a nominal variable by collapsing the values into categories. By specifying Years a Member Category as a nominal variable in the binary logistic regression function, we get the relevant output beginning on the next page.

The first thing presented is the coding that SPSS has generated, creating a new dummy variable for each value of the variable. Membership in a particular group is designated with a 1, and non-membership is designated with a zero. For example, individuals who have been a member for 3-5 years are coded with a 1 in group (1) and all other participants are coded with a 0 in this group. Also, members of the 3-5 year group are coded as a 0 in all other groups.

In this example, the first group (0-2 years a member) has no 1's. That is because this group is the reference group. SPSS gives users the option of making the first group in a nominal variable as the reference group, or the last group as the reference group.

Since all values of this variable are treated as a separate variable, this model will contain a unique B value for each group, and this B value will be that group's change in the logit *compared to the reference group*, instead of the incremental change in the logit based on number of years that we saw in the previous model.

This is the way that SPSS treats all nominal variables with more than two values (e.g. race, geographic region, formal education). Nominal variables with only two values (e.g. have a college degree – yes/no) can be treated as a continuous variable with a negative relationship for those who said no and a positive relationship for those who said yes.

Categorical Variables Codings

			Parameter coding					
		Frequency	(1)	(2)	(3)	(4)	(5)	(6)
Years a Member Category	0-2	506	.000	.000	.000	.000	.000	.000
	3-5	267	1.000	.000	.000	.000	.000	.000
	6-10	425	.000	1.000	.000	.000	.000	.000
	11-15	247	.000	.000	1.000	.000	.000	.000
	16-20	180	.000	.000	.000	1.000	.000	.000
	21-30	185	.000	.000	.000	.000	1.000	.000
	31 +	22	.000	.000	.000	.000	.000	1.000

Notice that the Nagelkerke R^2 increased from .224 when we treated Years a Member as a continuous variable to .237 when it is treated as a nominal variable, indicating a slightly better fit for the model.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1900.465[a]	.167	.237

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The accuracy of the model is statistically unchanged from 74.0% correct using the continuous variable, to 73.9% correct using the nominal variable.

Observed		Predicted Convention Attendance		
		Did not attend a Convention	Attended a Convention	Percentage Correct
Convention Attendance	Did not attend a Convention	1125	159	87.6
	Attended a Convention	320	228	41.6
Overall Percentage				73.9

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Years a Member Category			253.530	6	.000	
	Years a Member Category(1)	1.331	.227	34.456	1	.000	3.784
	Years a Member Category(2)	1.648	.204	65.506	1	.000	5.196
	Years a Member Category(3)	2.234	.216	107.205	1	.000	9.338
	Years a Member Category(4)	2.569	.228	126.629	1	.000	13.056
	Years a Member Category(5)	3.279	.233	197.732	1	.000	26.539
	Years a Member Category(6)	3.129	.476	43.253	1	.000	22.847
	Constant	-2.569	.173	220.724	1	.000	.077

a. Variable(s) entered on step 1: Years a Member Category.

Since there is no B value for individuals who have been a member for 0-2 years, calculating their probability of attending the annual convention uses only the constant.

(6) $e^{-2.569} = .0766$

(7) .0766 / (1 + .0766) = 7.1%

The probability of attending the annual convention for those who have been members for 3-5 years (group [1]) would be:

(8) 1.331 + (- 2.569) = - 1.238

(9) $e^{-1.238} = .28996$

(10) .28996 / (1 + .28996) = 22.5%

The benefit of running the procedure with a continuous variable as a categorical variable is to not assume that the B value holds throughout the entire range of values. By examining the group B values on the previous page, you can readily see that the most tenured members (31 + years) have a smaller B value (3.129) compared to those who have been a member for 21-30 years (B = 3.279).

Odds Ratio

In addition to calculating probabilities, SPSS also generates the odds ratio, and it can be found under the column heading Exp(B). The odds ratio is the change in the odds of being in the outcome group (e.g. convention attendance) with an incremental change in the independent variable.

In the last table on the previous page the odds ratio for group (1) – 3-5 years a member, is 3.784. Because we are using the independent variable in this model as a nominal variable with the first group as the reference group, it means that those who have been a member for 3-5 years are 3.784 times more likely to attend the annual convention than those who are in the 0-2 years group.

Earlier in this paper when the model was run with Years a Member as a continuous variable, the odds ratio was 1.123. Because it is a number greater than one, that means that with every increase in one year of membership, the odds of attending the annual convention increase by 1.123. For independent variables with a negative (inverse) relationship with the dependent variable, the B value will be negative, and the value for Exp(B) will be less than one – indicating that for an incremental increase in the value of the independent variable, the odds of the expected outcome go down.

Full Model

Once all of the relevant predictor variables have been identified, they can all be entered into the binary logistic regression model to arrive at the full predictive model. With the four independent variables included in the new model, the Nagelkerke R^2 is up to 46.2% and the model has classified 82% of cases correctly.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1514.316[a]	.325	.462

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Observed		Predicted Convention Attendance		
		Did not attend a Convention	Attended a Convention	Percentage Correct
Convention Attendance	Did not attend a Convention	1173	111	91.4
	Attended a Convention	218	330	60.2
Overall Percentage				82.0

One of the options available in SPSS is to calculate the probability that each person will be in the outcome group (in this case, attended the annual convention), and that information will be saved in the database for each member. Additionally, SPSS can add a variable to show into which group the particular case has been placed, eliminating all of the manual calculations.

The user can choose to sort or select cases based on his/her criteria, using the probabilities of the full model as the basis for the selection, and develop targeted lists with an expected higher likelihood of success.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Years a Member Category			119.203	6	.000	
	Years a Member Category(1)	1.128	.253	19.879	1	.000	3.088
	Years a Member Category(2)	1.230	.230	28.692	1	.000	3.420
	Years a Member Category(3)	1.837	.244	56.611	1	.000	6.275
	Years a Member Category(4)	2.036	.261	60.719	1	.000	7.660
	Years a Member Category(5)	2.610	.267	95.484	1	.000	13.602
	Years a Member Category(6)	2.311	.553	17.463	1	.000	10.088
	Donate to the Foundation	.949	.165	33.070	1	.000	2.583
	Local Volunteer			146.573	2	.000	
	Local Volunteer(1)	-.892	.253	12.476	1	.000	.410
	Local Volunteer(2)	-2.264	.215	110.463	1	.000	.104
	National Volunteer			23.736	2	.000	
	National Volunteer(1)	-.514	.501	1.050	1	.305	.598
	National Volunteer(2)	-1.458	.451	10.466	1	.001	.233
	Constant	.640	.485	1.738	1	.187	1.896

a. Variable(s) entered on step 1: Years a Member Category, Donate to the Foundation, Local Volunteer, National Volunteer.

In the table on the preceding page, one can see that the continuous variable, Years a Member, has been treated as a categorical variable, that one's status as a national volunteer or a local volunteer has also been added as a categorical variable. Whether or not the member donated to the foundation has been added, but as a continuous variable since this variable only has two options (yes/no).

Each of the three categorical variables have a reference group. The two volunteer groups have negative B values, indicating that former volunteers and never volunteers are less likely to attend an annual convention than current volunteers. Those groups with negative B values also have odds ratios Exp(B) less than 1, indicating that their odds of attending the convention are less than the odds of those who are in the reference group.

In summary, the purpose of using binary logistic regression is to use available information (survey data, membership application data, attendance records, continuing education logs) and use that information to better identify individual members and corporate members who are more likely (or less likely) to engage with the organization in a variety of ways. The predictive ability of the independent variables is based on relationships among the variables and is not intended to suggest causality.

Supplemental Readings

Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X. (2013) *Applied Logistic Regression*, *Third Edition*, Wiley Inc., Hoboken, New Jersey.

Kleinbaum, D.G., and Klein, M. (2010) *Logistic Regression – A Self Learning Text, Third Edition*, Springer. New York.

Menard, S. (2002) *Applied Logistic Regression Analysis*, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-106, Thousand Oaks, CA: Sage.

Pallant, J. (2016) *SPSS Survival Manual*, *Sixth Edition*, Allen & Unwin, Sydney.

About the Author

Larry J. Seibert, Ph.D. is the founder and CEO of Association Metrics, Inc. He can be reached at larry@associationmetrics.com or 317.840.2303.