

Background: In our summer 1999 newsletter¹, we discussed a method for the analysis of preference data. The goal of that method, called Multivariate Preference Mapping, was to show how preference data can be used to derive maps of products, ideals and their variances so that the basis for preference decisions could be understood. In this report we look at the application of a widely used tool, logistic regression analysis, to model preferential choice data. The logistic model is a staple for modeling risk factors in the field of epidemiology. This model does not allow for probabilistic perceptual variability and covariance as seen in probabilistic unfolding models². Nevertheless, while there are drawbacks to the use of logistic regression as a model for preference data, it can be effectively applied under certain circumstances and it offers a relatively simple technique that lends itself to easy interpretation.

Scenario: Your company produces and markets chocolate milk. You would like to test your existing product against an experimental product produced using a new pasteurization process. In a national study, 100 consumers of chocolate milk evaluated variants of the two types of milk which differed in product age. Chocolate milk at four equally-spaced product ages, T_1 , T_2 , T_3 , and T_4 beyond the expiration date were tested against the product at expiration date. Preferences are expressed in Table 1. It is known that the two processes produce identical product at their expiration dates.

CP T_0 vs.	Alternative			
	CP T_1	CP T_2	CP T_3	CP T_4
# Consumers choosing CP T_0	58/100	65/100	72/100	78/100
EP T_0 vs.	Alternative			
	EP T_1	EP T_2	EP T_3	EP T_4
# Consumers choosing EP T_0	54/100	62/100	66/100	70/100

CP: Current Product
EP: Experimental Product

Table 1. Preference proportions for current and experimental products at four different ages based on a sample of 100 consumers.

Choice and the Logistic Model: In consumer theory, it is assumed that when faced with a set of alternatives, a consumer makes a choice that has the highest utility, a numerical measure of worth. One could think of a liking response as an expression of utility; the higher the liking, the higher the utility. According to Luce's choice rule³, the probability of making a particular choice is obtained from the utility for the item chosen divided by the sum of the utilities for all the alternatives in a given set. When there are two alternatives for a particular pair the preference for u_1 is $u_1/(u_1 + u_2)$, where u_1 and u_2 are the utilities of the two alternatives. If we assume that the utilities depend on time so that $u_i = \exp(\beta T_i)$ where u_i is the utility for the product at time T_i , then we can use logistic regression to estimate β . If the two samples at

time 0 were compared, then the preference proportion would be 0.5, as expected.

Logistic regression analysis is a maximum likelihood technique that models the probability of observing a specific outcome from among a collection of possible outcomes⁴. The model provides a predictive probability for each possible outcome in terms of the levels of one or more explanatory variables. The predictive probabilities are determined by the model's estimates of the coefficients of the explanatory variables.

One of the reasons that logistic regression is an attractive tool is the ease with which the results can be evaluated and interpreted. The magnitude of the coefficient for an attribute, such as time in this example, is a measure of the sensitivity of the model to that attribute: the larger the magnitude, the greater the sensitivity. Even more helpful is the odds ratio for that variable. The odds ratio is a measure of how much the odds of observing the outcome change for some fixed change in the level of an explanatory variable. Odds ratios greater than 1 indicate a greater likelihood of observing the outcome being predicted; those less than 1 indicate a lower likelihood. For example, if we found the odds ratio to be 1.5, this would be interpreted in the following way: respondents are 1.5 times (or 50%) more likely to select the product at expiration date for each unit increase in the time period T . Because odds ratios are log normally distributed, it is easy to perform hypothesis tests to check for the significant influence of an explanatory variable.

Explaining the Preference Data: The data from the preference testing of chocolate milk contains preference information on the age of each product beyond expiration date. Two logistic regression analyses were run, one for the current product and one for the experimental product. The dependent variable was preferential choice. The independent variable was the time beyond expiration date.

You find that the fitted model did not differ significantly from a perfect fit for both the current product ($p = 0.99$) and the experimental product ($p = 0.97$), indicating good fits to the data. Also, for both products, time beyond expiration date was a significant driver of preference ($p < 0.001$ in both cases).

The odds ratio for the current product was 1.37. This means that consumers are 37 percent more likely to choose the younger product for each unit increase in product age difference. The odds ratio for the experimental product was 1.25; indicating consumers are only 25 percent more likely to choose the younger product for each unit increase. Both odds ratios were significantly different from 1, and the difference between the β values was significant ($p = 0.033$). This suggests that the new pasteurization process has a significant effect on consumer choice of chocolate milk over

time. This is also evident in the graphs of the two utility functions as a function of time, shown in Figure 1. As time increases, the utility of both products decreases. The utility function for the experimental product appears on the top. It does not decrease as rapidly as the utility function for the

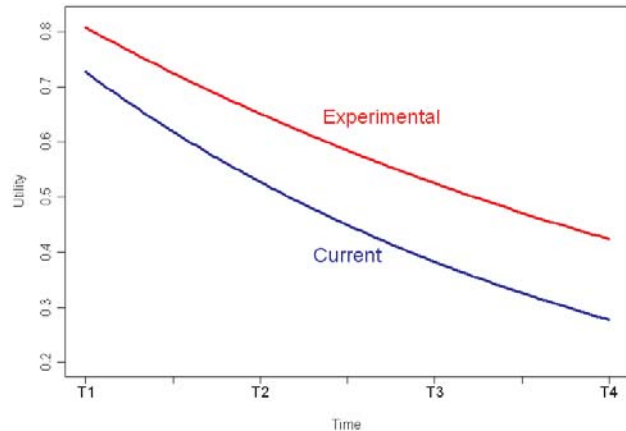


Figure 1. Utility functions for the current product and for the experimental product as a function of time.

Because the predictive choice probabilities depend on the time T_p , we can use the models to make estimates of the choice probabilities for time periods not included in the original study. For example, we could estimate the choice probability for the experimental product at a time midway between T_1 and T_2 , as the following table shows.

Observed			
EP T_0 vs.	EP T_1	EP $T_{1.5}$	EP T_2
# Consumers choosing EP T_0	54/100	---	62/100
Predicted			
EP T_0 vs.	EP T_1	EP $T_{1.5}$	EP T_2
# Consumers choosing EP T_0	55/100	58/100	61/100

Table 2. Observed and predicted choice probabilities from the model of consumer preference.

Analyses of this type can be useful if we are interested in estimating the point where respondents would choose the younger experimental product 60% of the time. In this case that estimate is $T_{1.9}$ and a 95% confidence interval for this estimate is $(T_{1.4}, T_{3.8})$. Comparisons among time periods can be made using ratios of the appropriate utility functions. For example, the predicted preference for the experimental product at time T_1 over the product at T_2 is 0.55.

Discussion: Logistic regression lends itself to a stepwise procedure in which explanatory variables are included in the model in order of their relative importance to the dependent variable. Unlike this example, there may be more than one

explanatory variable. The improvement to a model given by the inclusion of an additional explanatory variable can be measured by a test whose test statistic is Chi-square distributed. Stepwise logistic regression continues to add explanatory variables in order of importance until the test is no longer significant.

There are some potential problems to be aware of with the use of logistic regression. When modeling consumer choice using an explanatory variable that induces satiety such as sweetness level or bitterness, the model fit may not be acceptable. This is because logistic regression typically uses a linear exponential expression when a nonlinear expression would be required to effectively model the outcome. Other problems include those induced by sparse data⁵ and highly correlated explanatory variables. Sparse data occurs either when there is a small sample, or when a large number of explanatory variables are used relative to the sample size. Adequate cell counts (greater than 5) are necessary. Including explanatory variables that are highly correlated ($r > 0.60$) in a logistic model can lead to instability and unreliable odds ratios and utility functions. One of the two correlated explanatory variables should be removed from the model.

Conclusion: Logistic regression as a model of consumer preference can be effectively used to identify important attributes that drive the choice being modeled. The model can be used to tie measurable changes in product attributes to its performance in preferential choice. In more complex situations, more than one variable may be required to adequately model choice. In those circumstances, statistical tests on the odds ratios for these variables can in turn lead to an understanding of the hierarchy of their influence. Standard tests are available to assess model fit and model improvement as new variables are added. Variables known to induce satiety, such as sweetness or bitterness intensity, should be modeled with a nonlinear term. Care should be taken to avoid sparse data situations that can arise either from small sample sizes, from constructing a model with a large number of variables relative to the sample size, or from continuous explanatory variables. Preliminary analysis of explanatory variables should always be undertaken to eliminate one of any pair of highly correlated variables. In the end, the goal should be to construct the most parsimonious model possible that adequately explains the data.

References

1. Ennis, D.M. (1999). Multivariate preference mapping. *IFPress*, 2(2), 2-3.
2. Ennis, D.M. and Johnson, N.L. (1994). A general model for preferential and triadic choice in terms of central F distribution functions. *Psychometrika*, 59, 91-96.
3. Luce, R.D. (1959). *Individual choice behavior: a theoretical analysis*. New York: Wiley.
4. Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression*, (Second Edition). New York: Wiley.
5. McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models*, (Second Edition). New York: Chapman and Hall.