Evaluating Factors Influencing Grocery Store Choice

by
Marco Palma, Robert Emerson, and Lisa House [1]
July, 2003

Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Montreal, Canada, July 27-30, 2003

[1] Authors are research assistant, professor, and associate professor, respectively, Department of Food and Resource Economics, University of Florida.

Introduction

The food sector, in particular at the retail level in the food distribution chain, continues to have structural changes, with some of the changes being reflected by consumer choices regarding when and where to make food purchases. Supermarkets have grown substantially and become more concentrated, while there has been entry of new grocery store forms such as super-centers and large food distribution warehouses. These new options provide customers with alternatives that may increase competition (Medina and Ward, 1999).

In recent years, Wal-Mart has made significant impacts on the retailing business by combining general merchandise stores with full size supermarkets to form what is known as Wal-Mart Supercenters. The first Wal-Mart Supercenter opened its doors in Washington, Missouri in 1988. Eleven years after the opening of the first supercenter, the Wal-Mart Company was the fourth largest food retailer in the nation and expanded to international markets (Huang et al, 2002).

This paper analyzes consumer preferences toward grocery store choices given a set of attributes of stores. This information will then be used to make inferences on how the opening of a Wal-Mart supercenter would affect the other grocery stores in a small city.

Background

The number of non-traditional grocery outlets has increased substantially in the past few years. Nontraditional outlets target specific high volume categories of dry grocery products, paper products, frozen foods, limited perishable produce and meat products, health and personal care products, and general merchandise. Low operating margins provide attractive low priced products to consumers while ensuring high volume shipments by suppliers (Capps and Griffin, 1998).

In this study, residents of a small city (population approximately 22,000 residents), were surveyed about their choice of grocery stores. At the time of the survey, three grocery stores existed in the city, with residents aware a Wal-Mart Superstore was opening within two months. The three existing grocery stores represented three different types of grocery stores: Store A (generally regarded as higher quality, higher price); Store B (generally regarded as medium quality, medium price); and Store C (generally regarded as low-quality, low price). All three grocery stores and the planned Wal-Mart Supercenter were located within six miles on the same main thoroughfare.

Data and Methods

The data for this study was obtained through a mail survey. Four hundred surveys were mailed to randomly selected households (fitting the criteria of having children in the house and incomes above $25,000). An effective response rate of 18.75% (75 surveys) was achieved. The survey included questions ranking the three grocery stores currently in the market on quality attributes and price; which store the respondent shopped at for major and minor purchases; which store they were located closest to; and what the top three reasons were for choosing a grocery store. Demographics were also collected. The demographic questions included number of people in the household, age, gender, level of education, income, etc.

A discrete choice framework is used to analyze respondent preferences for a grocery store. Although the questionnaire asked for preferences among three different stores, the third store had an inadequate number of complete observations (5) for analysis. The dependent variable representing the preferred grocery store is therefore a binary variable. Among available explanatory variables are demographic variables, respondent rankings of the most important qualities in a grocery store, and likert scale variables representing the respondents' evaluations of each of the existing grocery stores.

The basic starting point for unordered choice models is typically random utility theory. Given two alternatives, the utility for the choice 1 may be represented for the i[th] consumer as:

$$U_{i1} = \beta' z_{i1} + \varepsilon_{i1} \tag{1}$$

If the consumer chooses alternative 1, then it is assumed that the U_{i1} is the larger of the 2 utilities. Then the statistical model is driven by the probability that choice 1 is made, which is (Greene, 2000):

$$\Pr(U_{i1} > U_{i2}). \tag{2}$$

The available data permit the estimation of alternative forms of the discrete choice model. The basic logit model explains the consumer's choice on the basis of characteristics of the individual decision-maker. The existing data set includes various demographic variables for the individual as well as the individual's top three reasons for their choice of store. In this case, the logit estimation is based on the following specification of the probabilities:

$$\Pr(y = 1 \mid x) = \frac{\exp\left(\sum_{k=1}^{K} \alpha_{1k} x_k\right)}{\left[\sum_{j=1}^{2} \exp\left(\sum_{k=1}^{K} \alpha_{jk} x_k\right)\right]} \tag{3}$$

The explanatory variables x_k in equation (3) include the individual demographic characteristics and their top three reasons for choice of store . Although equation 3 indicates two different sets of α parameters to be estimated, only one set can be estimated since they must always be normalized relative to a base set; in a model with *m* choices, only *m-1* independent sets of

Comment: Probably needs to be taken out

parameters can be estimated. The questions that can be answered with this type of specification are what types of individuals are most likely to select one store over another based on their characteristics. For example, does a particular store appeal to low-income or high-income consumers? An important aspect of the approach is that can only answer questions pertaining to the stores under consideration in the sample.

An alternative approach is one that is based on McFadden's conditional logit model. The conditional logit model utilizes the individual's evaluation of specific attributes of each of the alternatives under consideration in explaining the discrete choice decision. In this case, the probability model is specified as:

$$\Pr(y_i = 1 | z) = \frac{\exp(\beta' z_{i1})}{\sum_{j=1}^{2} \exp(\beta' z_{ij})} \qquad (4)$$

There are K z_{ij} variables, each representing the evaluation of the k^{th} attribute of the j^{th} choice by the i^{th} individual. Note that there is only one set of β parameter estimates for this specification; they are not specific to the choices considered. Among the attributes available in the data set for inclusion in this specification are each respondent's evaluation via a likert scale of a set of grocery store attributes for each store. An indication of which store is closest for them is also available and suitable in this specification. An important characteristic of this specification is the ability to make predictions about store choice for a store *not* in the sample. Given an evaluation of attributes of a new store not in the sample, a prediction may be made about the probability of selection of the new store. This is not possible with the basic logit model specified earlier.

A third form of logit model combines the above two models, including both individual characteristics and evaluations of attributes of the choices. This form of the model permits both types of predictive statements: which type of consumer is most likely to select a particular store among those in the sample, but in addition, permits an evaluation of the probability of selection of a new store not in the sample based on an evaluation of the attributes of the new store by an individual. The specification of the probabilities is:

$$\Pr(y_i = 1 \mid x, z) = \frac{\exp(\beta' z_{i1} + \alpha_1' x_i)}{\sum_{m=1}^{2} \exp(\beta' z_{im} + \alpha_m' x_i)} \qquad (5)$$

The different forms of the logit model are estimated using standard maximum likelihood techniques.

Results And Discussion

Estimates of the standard logit model are reported in Table 1. Among the respondents, 42 selected the Store A store and 28 selected Store B as their preferred store. The average number of people in the household was 3.67. About half of the respondents preferred the grocery store closest to them. The average age of the respondents was 45 years.

The log likelihood function for the estimated model had a value of –29.30504; the restricted log likelihood function (no explanatory variables included) was –47.11082. The resulting estimated model is highly significant with a likelihood ratio test statistic value of 35.61156. With 19 degrees of freedom, the null hypothesis of zero values for the parameters is rejected at better than the 5% level.

These results from the binomial logit model reveal that three characteristics are statistically significant in determining store choice: price, variety and closest store. The results for the demographic characteristics were disappointingly statistically insignificant. The general manager of the Store A store identified one of the same characteristics that is indicated as discriminating between the two stores: variety of product.

Among the 70 complete observations used in the model, 42 reported Store A as the preferred grocery store and 28 reported Store B as the preferred grocery store. Table 2 presents the frequencies of actual and predicted outcomes. Despite the limited statistical significance of individual parameters, the model predicts 53 of 70, or 75.71 percent, of the observations correctly. This percentage was obtained by adding up the diagonal elements (34+19) and dividing by the number of observations (70) in table 2. A naïve model which always predicts $y=0$ because $P<0.5$, predicts 42 of 70, or 60 percent of the observations correctly.

The conditional logit estimates are displayed in table 3. Responses were available for the following 12 attributes of the stores as evaluated by the respondents: service, quality, price, variety, convenience, friendliness, knowledge of staff, location, cleanliness, lighting, layout, and return policy. Each was rated by the respondents on a scale of one to five where one represents "needs improvement," three is "satisfactory," and five is "excellent." Although a number of the questions appear the same as the questions used in the standard logit, there is a significant difference: the questions for the standard logit were a 1-3 ranking of the *selected* store; the responses used for the conditional logit are a rating of *each* attribute for *each* store by *each* person.

Since a number of the attributes appear fairly similar, a principle components analysis of the attributes was done to determine the extent to which they are measuring the same characteristics. Since the first five principle components represented over 85 percent of the variation, five of the attribute variables were selected to avoid a lack of identification of any of the attributes. The ones selected were on the basis of those most likely to have consistent perceived benefits to the consumer: price, variety, quality, cleanliness, and service. In addition, whether or not the store was closest for the consumer was a choice based attribute as well. With this set of variables, there are 61 complete observations.

The conditional logit estimates in table 3 suggest that the closeness of the store is one of the stronger variables. The partial effect (evaluated at the means) suggests that the probability of shopping at the closer store is 0.13 larger than at the more distant store. In addition, price and service are at the margin of standard significance levels. Better price performance and better service as evaluated by the consumer also increase the probability of shopping at the store performing better on those attributes.

The mixed logit estimates are also displayed in table 3. While including the same attribute variables as the conditional logit, the specification also includes the individual characteristics of the respondents. The comparison of the two specifications yields a likelihood ratio statistic of 23.13. With nine degrees of freedom, the individual characteristics are jointly significant at better than the one percent level. Among the attributes, closeness of the store remains strongly significant. While price performance is only marginally significant, quality perception is strongly significant. All three of the significant attribute variables have a positive effect on store selection. Cleanliness and variety do not appear to be determining factors in store choice, at least between the two stores under consideration.

Among the individual characteristics, household size is statistically significant as is being in the $60,000 - $99,999 income group. Given the normalization, the positive parameter for household size implies that the larger the household size, the less likely is the individual to shop at Store A, or conversely, the more likely they are to shop at Store B. Since the parameter on the significant income group is negative, the upper middle-income group has a higher probability of choosing Store A for grocery shopping. Education, gender, and age appear to have little systematic effect on store choice.

In contrasting the estimates for the three specifications, the mixed logit model appears to have some advantages in interpreting the data. It is able to identify the effects of some of the individual characteristics which the standard logit did not. In addition, it is no weaker in identifying the significant attributes for consumers (recall that the measurement of the attributes differs in the two cases). The straight conditional logit handles the attributes well, but the mixed logit specification appears to represent a modest improvement.

Concluding Remarks

Traditional grocery stores face stronger competition from mass merchants, such as Wal-Mart. In order to maintain market share traditional grocery stores must identify a way to compete. Information on what factors influence a customer to choose a grocery store can aid in understanding customer reaction to new competitors, including new competitors such as Wal-Mart. According to the conditional and mixed logit models, the characteristics that have a statistically significant influence on grocery store choice are price and location of the grocery store (distance). In the mixed model, quality was significant, and in the conditional model, service marginally significant. The standard logit model isolated price, variety and location. The two variables consistently significant across the different specifications are price performance and location.

Since one of Wal-Mart's prime characteristics is perceived price performance, the estimation results suggest that consumers will respond and existing stores are likely to face a loss of customers. Additionally, Store A can expect a bigger impact in the study as it was located near the Wal-Mart. In other words, respondents who were located closest to Store B would remain located closest to Store B, but respondents who were closest to Store A now could be closest to Wal-Mart.

Some of the questions incorporated within the market survey included if customers had shopped at a Wal-Mart super-center before, if they plan to shop at the new Wal-Mart super-center. Most (91 percent) of the respondents have shopped at a Wal-Mart super-center before, and 100 percent plan to shop at the Wal-Mart super-center. These responses in conjunction with the results from the model presented in this paper suggest that Wal-Mart will be a strong competitor for the existing stores. Finally, the recommendation to the grocery stores analyzed in this paper (i.e. Store A and Store B) is to emphasize price performance.

References

Capps, O. and J. M. Griffin. "Effects of a Mass Merchandiser on Traditional Food Retailers". *Journal of Food Distribution Research* 29(1):1-7.

Greene, W.H. *Econometric Analysis,* Fourth Edition. Prentice Hall, Upper Saddle River, New Jersey, 2000.

Greene, W.H. *Limdep Version 7.0 User's Manual Revised Edition.* Econometric Software, Plainview, New York, 1998.

Huang, L.C., J.E. Epperson, B.J. Cude, and B. Woo. "Wal-Mart Supercenter The New Low Price Food Retailer in Town." *Choices* Summer 2002:6-9.

McFadden, D. Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka, ed., *Frontiers in Econometrics*. New York: Academic Press, 1973.

Medina, S. and R.W. Ward. "A Model of Retail Outlet Selection for Beef." *International Food and Agribusiness Management Review*, 2(2)1999: 195-219.

Table 1. Standard logit model estimates of store choice ($Pr[y_i$=Store B])

Variable	Coefficient	Estimated Standard Error
Constant	-4.2287	2.9634
Service[a]	-0.2642	0.4676
Price[a]	0.7002	0.4020
Variety[a]	0.9463	0.4632
Convenience[a]	-0.2101	0.4064
Location[a]	0.2553	0.3995
Atmosphere[a]	0.1813	0.3692
High quality produce[a]	0.3929	0.5129
High quality meat[a]	0.0994	0.4645
Overall quality[a]	-0.2045	0.3490
Knowledgeable staff[a]	-0.8527	0.7925
Closest store	2.9560	1.0749
Household size	-0.7012	3.9393
Age	0.3791	0.4535
Gender	-0.9326	0.9411
College education	1.5600	1.0886
Some Graduate School	-0.4242	1.09612
Income: $30,000 - 59,999	0.3527	0.9610
Income: $60,000 – 99,999	-1.019	1.2109
Income: >$99,999	-1.5258	1.4915

Log Likelihood: -29.3050

Observations: 70

[a]Variable was given a rank of 1, 2 or 3 by the respondent as most important reason for selecting the store.

Table 2. Frequencies of Actual and Predicted Outcomes

	Predicted		
Actual	0	1	**Total**
0	34	8	**42**
1	9	19	**28**
Total	**43**	**27**	**70**

Table 3. Conditional logit and mixed logit estimates

	Variable	Model					
		Conditional Logit			Mixed Logit		
		Coefficient	**Std. Error**	**∂Pr(A)/∂x**	**Coefficient**	**Std. Error**	**∂Pr(A)/∂x**
Store Attributes	Closest Store	1.1469	0.4582	0.1347	4.8839	2.1402	0.2709
	Price	0.8155	0.4374	0.0958	1.3032	0.8254	0.0723
	Variety	0.4364	0.5753	0.0513	-1.8651	1.2963	-0.1034
	Quality	0.7106	0.6525	0.0835	4.2483	2.0541	0.2356
	Cleanliness	0.0953	0.5478	0.0112	-0.8880	0.9861	-0.0492
	Service	0.6277	0.3913	0.0737	1.1681	0.8479	0.0648
Individual Characteristics	Constant				-8.7006	4.9397	0.4825
	Household size				1.8527	0.8623	-0.1027
	Age				0.1111	0.0760	-0.0062
	Gender				0.0486	1.4347	-0.0027
	College Edcn				-1.1800	2.0857	0.1043
	Graduate School				-4.6468	3.0813	0.2577
	Income: $30,000-59,999				4.9487	3.4510	-0.2744
	Income: $60,000-99,999				-5.7739	2.5830	0.3202
	Income: >$99,999				-3.0425	3.0961	0.1687
	Log Likelihood		-22.6007			-11.0356	
	Observations		61			61	
	Choices		2			2	