Analysis of Cross Category Dependence in Market Basket Selection

GARY J. RUSSELL
University of Iowa

ANN PETERSEN
Drake University

Market basket choice is a decision process in which a consumer selects items from a number of product categories on the same shopping trip. The key feature of market basket choice is the interdependence in demand relationships across the items in the final basket. This research develops a new approach to the specification of market basket models that allows a choice model for a basket of goods to be constructed using a set of “local” conditional choice models corresponding to each item in the basket. The approach yields a parsimonious market basket model that allows for any type of demand relationship across product categories (complementarity, independence, or substitution) and can be estimated using simple modifications of standard multinomial logit software. We analyze the choice of four grocery store categories that exhibit common cross-category brand names for both national brands and private labels. Results indicate that cross-category price elasticities are small. We argue that store traffic patterns may be more important than consumer-level demand interdependence in forecasting market basket choice.

The advent of retail scanner data has created a wealth of information about consumer behavior for both retailers and manufacturers. This information revolution has allowed researchers to study the pattern of brand competition within specific product categories (e.g., Cooper, 1988; Grover and Srinivasan, 1992; Russell and Kamakura, 1994). However, researchers and managers are increasingly interested in understanding the pattern of brand competition across product categories (Blattberg, 1989, Blattberg and Neslin, 1990).

Gary J. Russell is Professor of Marketing at the Tippie College of Business, University of Iowa, Iowa City, Iowa 52242-1000, (e-mail: gary-j-russell@uiowa.edu). Ann Petersen is Visiting Assistant Professor, College of Business, Drake University, Des Moines, Iowa 50311-4505, (email: apetersen@blue.wecg.uiowa.edu).

Journal of Retailing, Volume 76(3) pp. 367–392, ISSN: 0022-4359
Copyright © 2000 by New York University. All rights of reproduction in any form reserved.
MARKET BASKET ANALYSIS

Our research is based upon a growing body of work known as market basket analysis. Market basket analysis is a generic term for methodologies that study the composition of the basket (or bundle) of products purchased by a household during a single shopping occasion. Current commercial applications emphasize electronic couponing, the tailoring of coupon face value and distribution timing using information about the household’s basket of purchases (Catalina Marketing, 1997); and affinity analysis, the design of store layout according to the coincidence of pairs of items in a market basket (Brand and Gerristen, 1998). Both types of applications are based upon the belief that sales in different product categories in the market basket are correlated. The patterns in these correlations are then used to make marketing strategy recommendations.

The academic literature in market basket analysis is small, but growing. This literature attempts to go beyond the correlational approaches found in the marketing research industry by identifying the sources of cross-category dependence in market basket selection. Market basket analysis is regarded as a pick-any choice problem. When consumers enter a store, they are confronted by a large number of possible product categories that may be purchased. Consumers may then select all, none, or any subset of the available categories. The key questions facing choice researchers in marketing are whether the multiple decisions in a pick-any choice task are related and how marketing managers can use cross-category linkages to develop marketing strategies (Russell, Ratneshwar, Shocker et al., 1999).

Two different explanations for cross-category linkage have been suggested: store choice and global utility. Store choice models argue that sales in different categories are related because the mix of consumers in the store changes from week to week due to marketing activity. Because product preferences are correlated across categories (Russell and Kamakura, 1997), the changing mix of consumers over time creates cross-category correlation in store-level sales data.

For example, Bodapati and Srinivasan (1999) develop a model in which consumers use knowledge about prices and feature advertising to select a retailer who delivers the least expensive basket of goods. Once the retailer is selected, the basket is formed by a series of conditionally independent consumer decision models predicting category incidence, brand choice and purchase quantity. Bell and Lattin (1998) develop a model that relates store choice to the pricing strategy of a store across multiple product categories. The decision to make a major shopping trip (i.e., allocate a larger dollar expenditure to the basket) alters store choice and also increases the number of items in the market basket. Models in this research stream assume that cross-category dependence is found at the market level (due to store traffic effects), but is not found at the consumer level (because the household does not view purchases in different categories as related).

In contrast, global utility models argue that cross-category dependence is present within the choice process of each consumer. In these models, cross-category choice correlations exist because consumer preference for an item in one category is contingent on the consumption of items in other product categories. Harlam and Lodish (1995) link choices across potential complements within the same product category (different flavors of
powdered soft drinks) by making utility for the current choice dependent upon the
attributes of previously selected items (the flavors of previously selected drinks). Erdem
(1998) demonstrates that the utilities for products in different product categories that share
the same brand name jointly covary with product experience. Manchanda, Ansari and
Gupta (1999) use the multivariate probit model to show that true consumption comple-
ments (detergent and fabric softener) exhibit choice dependence within a shopping trip. In
effect, global utility models argue that cross-category choice dependence exists even when
store traffic effects are unimportant.

RESEARCH AGENDA

Our work extends the market basket literature by proposing a new method of studying the
strength of cross-category choice dependence within each consumer’s purchase history.
Drawing upon modeling techniques from the spatial statistics literature, we show how a
researcher can develop a global utility model for the entire basket of purchases by
constructing linked choice models for each category individually. This approach has two
important advantages. First, we obtain a parsimonious global utility model that allows for
any type of demand relationship across product categories (complementarity, independ-
ence, or substitution) within the choice process of each consumer. That is, the within-
consumer pattern of demand effects is unrestricted. Second, because the building blocks
of the approach have a logistic form, we demonstrate that the resulting market basket
choice probabilities follow a multivariate logistic distribution. As we show subsequently,
this model form allows response parameters to be estimated using simple modifications of
standard multinomial logit software.

We begin by discussing a general method of building a complex global choice model
for the market basket from relatively simple single category choice models. By special-
izing this general approach, we derive the multivariate logistic market basket model, and
explore its implications for consumer choice behavior. We then apply the approach to the
choice of four grocery store categories that exhibit common cross-category brand names
for both national brands and private labels. Although we find evidence of cross-category
choice dependence, the magnitude of the cross-category price effects is modest. Substan-
tively, our results suggest that store traffic effects may be more important than global
utility effects in modeling market basket choice.

BUILDING A MARKET BASKET CHOICE MODEL

The market basket model developed in our research is built using a flexible approach to
model specification developed in the spatial statistics literature. This approach, which we
call conditional choice specification, essentially allows the researcher to specify a global
model (the choice of entire basket of items) by specifying a series of local models (a
choice model for each item in the basket). Intuitively, this method assumes that choices
are made in a certain order, but does not require the researcher to actually know the order in which choices are made in building up the basket.

Multiple Category Choice

Suppose that the researcher is attempting to model choice activity across four product categories: A, B, C, and D. Consumers make choices in each of the four categories in some sequence that is not observed. In a statistical sense, we can think of each basket as consisting of four random variables (corresponding to the buy and not buy decisions of the consumer). Clearly, the choice process implies that the consumer will select one of \(2^4 = 16\) possible market bundles. This is a pick-any choice task: consumers may select all, none, or any subset of the four available categories.

Formally, the choice process for the entire basket can be expressed in terms of a four dimensional multivariate distribution \(p(A,B,C,D)\) that defines the relative likelihood of each of the 16 possible market baskets. At this point, two different strategies may be advanced for constructing a choice model. First, we could attempt to directly specify \(p(A,B,C,D)\) based upon some prior knowledge of how category features interact in the consumer’s utility function. This is equivalent to a direct specification of a global utility function over a bundle of items [see, e.g., Farquhar and Rao (1976) and McAlister (1979)]. Second, we could assume that the choices that lead to the construction of a basket are made in a known order [Kamakura et al. (1991), Harlam and Lodish (1995)]. Using this sequential information, we could then develop a model in which the consumer evaluates the utility of the current item relative to a cumulative variable representing the composite utility of the basket of previous choices. For example, if D is considered first, C second, B third, and A fourth, then we would write:

\[
p(A,B,C,D) = p(A|B,C,D) \cdot p(B|C,D) \cdot p(C|D) \cdot p(D)
\]

where the notation \(p(x|y)\) denotes the probability of \(x\) given \(y\).

Both approaches are difficult to implement in a retail market basket setting. Directly specifying the utility function for the market basket is problematic because it requires both a detailed understanding of cross-category demand relationships (complementarity, independence, or substitution) and an explicit enumeration of all possible market baskets. In contrast, building the utility model sequentially is attractive conceptually, but makes the unrealistic assumption that the retailer can readily observe the order in which choices are made.

Conditional Choice Specification

The conditional choice specification approach proposed here avoids both the direct specification of the joint distribution \(p(A,B,C,D)\) and the assumption of a particular
choice order. The logic of the approach is best explained by considering the following scenario. Suppose that we follow the consumer around the store and observe each choice decision. Towards the end of the shopping trip, we find that the consumer has made choices in three categories (A, B, and C) and is now considering whether or not to buy in the last category (D). The conditional distribution approach assumes that we can specify \( p(D|A,B,C) \), the probability of buying in this last category, given the known outcomes of the previous choice decisions. In consumer behavior terms, specifying this conditional distribution is equivalent to specifying the utility of category D given the attributes of category D and the context created by the earlier choices.

However, recall that the researcher does not usually know the identity of the last category. For this reason, we replace assumptions about sequence with assumptions about the following set of conditional distributions: \( p(A|B,C,D) \), \( p(B|A,C,D) \), \( p(C|A,B,D) \), and \( p(D|A,B,C) \). These so-called full conditional distributions correspond to placing each category (in turn) as the last choice decision and then specifying the conditional probability of selecting this last category. Although the true decision sequence is not known, the researcher is assumed to be able to develop a set of choice models that collectively describe the last decision in any possible decision sequence.

Remarkably, by using only information about these full conditional distributions, the researcher can infer all properties of \( p(A,B,C,D) \), the probability distribution that describes the relative likelihood of each possible bundle. Technically, this can be done because there exists a one-to-one correspondence between full conditional and joint distributions: for any probability model (regardless of structure), the complete set of full conditional distributions uniquely determines the joint distribution \( p(A,B,C,D) \) provided that all full conditional distributions are mutually consistent (Besag, 1974; Cressie, 1993). Details on this theoretical relationship are discussed in the next section. Intuitively, the procedure works because the full conditional distributions completely define the dependencies in choice decisions across the entire choice bundle \( \{A,B,C,D\} \).

The conditional choice approach is of great practical importance because it is much easier for a researcher to use marketing theory to specify a choice process one decision at a time (the full conditional distributions) than to specify a choice process for the entire market basket simultaneously (the joint distribution). By focusing attention on only one choice decision, the researcher can draw upon the marketing literature on single category choice (e.g., Kamakura and Russell, 1989; Gupta, 1988; Jain, Vilcassim, and Chintagunta, 1994) during model specification. Nevertheless, the result of the conditional choice approach is a complete market basket model that implies dependence of choice on decision sequence and allows for flexibility in demand relationships among categories (complementarity, independence, or substitution).

**A Multivariate Logistic Model of Market Basket Selection**

In this section, we construct a market basket model by assuming that the conditional probability of choice in one category, given the actual choices in all other categories, can be expressed in the form of the logit model. We demonstrate that the implied choice model
for all market baskets can be expressed as a multivariate logistic distribution (Cox, 1972). This model permits the researcher considerable flexibility in assessing the degree of correlation among choices in different categories. In particular, it allows the researcher to predict how marketing activity in one category impacts choice in other product categories.

**Choice Problem**

Let $k$ denote a consumer and $t$ denote a time point. Assume that the consumer has $i = 1, 2, \ldots, N$ categories available for purchase. Then, we define a market basket as the vector of category choices

$$B(k,t) = \{C(1,k,t), \ldots, C(N,k,t)\}$$

where $C(i,k,t) = 1$ if consumer $k$ buys category $i$ at time $t$ (and equals 0 otherwise). Because each $C(i,k,t)$ can take on only two values, our notation implies that there are $2^N$ possible baskets that could be selected. (One of these baskets—the null basket—corresponds to nonpurchase across all categories.) Accordingly, the market basket model developed subsequently assigns a choice probability to each of these $2^N$ baskets.

**Conditional Choice Models**

The conditional choice methodology requires the researcher to specify the probability that each category will be chosen, conditional upon the known choices in all other categories. Here, we assume that the conditional utility of consumer $k$ for product category $i$ at time $t$ is given by

$$U(i,k,t) = \beta_i + HH_{ikt} + MIX_{ikt} + \sum_{j \neq i} \theta_{ijk} C(j,k,t) + \epsilon(i,k,t)$$

(2)

where $HH_{ikt}$ denotes variables defining household characteristics, $MIX_{ikt}$ denotes variables defining the marketing mix variables, and $\epsilon(i,k,t)$ is a random error with mean zero. The term $\sum_{j \neq i} \theta_{ijk} C(j,k,t)$ links the choice of the current category $i$ to the actual choice decisions in all other product categories in the basket.

Note that $\theta_{ijk} > 0$ implies a positive association between product categories, while $\theta_{ijk} < 0$ implies a negative association. For reasons which will become clear subsequently, logical consistency requires that the coefficients on the observed choice variables be symmetric ($\theta_{ijk} = \theta_{jik}$). Because $\theta_{ijk}$ depends upon the household $k$, we allow the magnitude of the cross effects to vary across households.

Without loss of generality, we assume that the probability of buying category $i$, conditional upon the choice outcomes in all other categories, equals the probability that $U(i,k,t) > 0$. Further, by assuming that the random error has an extreme value distribution, the conditional probability of selecting category $i$ can be expressed as the logit model
Cross Category Dependence

\[ \Pr(C(i,k,t) = 1|C(j,k,t) \text{ for } j \neq i) = [1 + \exp\{-Z(i,k,t)\}]^{-1} \quad (3) \]

where

\[ Z(i,k,t) = \beta_i + HH_{ik} + MIX_{ik} + \sum_{j \neq i} \theta_{ij} C(j,k,t) \quad (4) \]

is the deterministic portion of equation (2). Intuitively, one assumes in this model that the consumer’s choice of the final category in the basket (category i) is affected by the bundle of categories already selected. In this way, the probability of choice in one category is dependent upon the context created by previous choices.

Market Basket Model

Although the full conditional models in Equations (3) and (4) implicitly link all categories in a common framework, the form of the market basket distribution is not evident by simple inspection. For this task, we turn to Besag’s (1974) remarkable characterization theorem, which provides a simple mathematical way of deriving a joint distribution given a set of full conditionals (see Appendix A). In this application, the set of full conditional distributions is given by Equations (3) and (4), whereas the joint distribution refers to the distribution of the market baskets \( B(k,t) = \{C(1,k,t), \ldots, C(N,k,t)\} \).

By applying Besag’s (1974) theorem, we obtain the following key result. Suppose that basket b has contents \( \{X(1,b), X(2,b), \ldots, X(N,b)\} \). In this notation, \( X(i,b) \) is a dummy (0–1) variable that takes on the value one if category i is included in basket b. Then, given Equations (3) and (4) and the assumption that cross effect coefficients are symmetric \( \theta_{ik} = \theta_{ik} \), the probability of selecting basket b is given by

\[ \Pr(B(k,t) = b) = \exp\{\mu(b,k,t)/\sum_b \exp\{\mu(b^*,k,t)\}\} \quad (5) \]

where

\[ \mu(b,k,t) = \sum_i \beta_i X(i,b) + \sum_i HH_{ik} X(i,b) + \sum_i MIX_{ik} X(i,b) + \sum_{i \neq j} \theta_{ij} X(i,b)X(j,b) \quad (6) \]

is the imputed utility of basket b. Notice that this model predicts the probability of selecting each of the \( 2^N \) baskets using only the set of parameters needed to define the conditional logit models.

An appreciation of the structure of the model can be obtained by considering a simple setting in which a given consumer k chooses from only two categories (X1 and X2). In Table 1, we show the terms for \( \mu(b,k,t) \) corresponding to the four possible market baskets. Notice that the structure the \( \mu(b,k,t) \) is identical to the terms of a log linear model.
TABLE 1

<table>
<thead>
<tr>
<th>Structure of Market Basket Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Values of $\mu(b,k,t)$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Category 1</td>
</tr>
<tr>
<td>Present in Basket (X1 = 1)</td>
</tr>
<tr>
<td>$\beta_1 + M_{X_1}$</td>
</tr>
<tr>
<td>$+ \beta_2 + M_{X_2}$</td>
</tr>
<tr>
<td>$+ \theta_{12}$</td>
</tr>
<tr>
<td>$\beta_1 + M_{X_1}$</td>
</tr>
<tr>
<td>Category 1 Absent from Basket (X1 = 0)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note: For expositional reasons, this table ignores consumer specific effects. The probability of selecting a given market basket $b$ equals $\exp[\mu(b,k,t)]/\Sigma_{b'} \exp[\mu(b',k,t)]$ where the summation runs over all possible baskets $b'$.

containing main effects and two way interactions. In particular, it is easy to show that the cross effect term obeys the relationship

$$\theta_{12} = \left( \frac{\Pr(X1 = 1, X2 = 1)/ \Pr(X1 = 0, X2 = 1)}{\Pr(X1 = 1, X2 = 0)/ \Pr(X1 = 0, X2 = 0)} \right)$$

(7)

where the right hand side is the so-called odds ratio measuring association in the table. Because the odds ratio is symmetrical in the category indices (1 and 2), the cross effect coefficient must be symmetrical as well. Intuitively, this explains why symmetry in the $\theta_{ijk}$ is necessary to derive the basket model. In general, $\theta_{ijk}$ has the same sign (positive, negative, or zero) as the correlation between C(i,k,t) and C(j,k,t).

Model Interpretation

This model can be interpreted in two different (but logically equivalent) ways. Viewed from the standpoint of the $2^N$ market baskets $b$, Equations (5) and (6) can be viewed as a logit choice model defined over a set of alternatives (baskets) with a particular utility specification $\mu(b,k,t)$. This interpretation facilitates model calibration since standard logit software can be easily adapted to obtain maximum likelihood estimates of the parameters in Equation (6). It is important to understand that the set of variables $\{X(i,b) \text{ for } i = 1, 2, \ldots, N\}$ simply describe the contents of each particular basket. For example, the null basket (no purchases in any category) has the following set: $X(i,b) = 0$ for all categories $i$. In contrast, a basket consisting only of category 1 has the following set: $X(1,b) = 1$ and $X(i,b) = 0$ for all categories $i$ different from category 1. Clearly, the $X(i,b)$ are known to the researcher because they define the types of baskets available to the consumer. For this reason, once the market basket model in Equations (5) and (6) is calibrated, it may be used for forecasting purposes.

The model can also be interpreted as a multivariate distribution defined over the vector
of binary random variables \( B(k,t) = \{C(1,k,t), \ldots, C(N,k,t)\} \). Equations (5) and (6) are in the form of the multivariate logistic distribution (Cox, 1972), a general distribution for correlated binary random variables. For this reason, it is equally valid to state that the joint probability \( \Pr(C(1,k,t) = X(1,b), C(2,k,t) = X(2,b), \ldots, C(N,k,t) = X(N,b)) \) is given by Equation (6). In particular, the conditional probabilities \( \Pr(C(i,k,t) = 1 \mid C(j,k,t), j \neq i) \) implied by Equation (6) are identical to the conditional logit expressions assumed in Equations (3) and (4). As we show in our empirical work, this view of the market basket model facilitates the computation of cross-category price elasticities.

This discussion points up an important fact about the conditional choice specification approach. The market basket model in Equations (5) and (6) is not derived from a standard random utility maximization argument. In particular, we do not begin with the utilities in Equation (6), add a random error, and then derive choice probabilities for the market baskets. Instead, we begin with the conditional logit models in Equations (3) and (4) and then derive the implied market basket model [Equations (5) and (6)] using Besag’s (1974) theorem. According to this theorem, the only market basket model consistent with the assumed conditional logit models is the multivariate logistic distribution of Cox (1972). Accordingly, the logit form of the basket model follows from our conditional choice distribution assumptions. Put another way, once a researcher accepts the form of the conditional choice models in (3) and (4), the multivariate logistic form of the market basket model follows immediately as a logical consequence.

Summary

At this point, the key features of the multivariate logistic basket model should be clear. By introducing cross effects into the conditional logit models of Equations (3) and (4), we are able to build a parsimonious basket selection model that accommodates a general pattern of dependence across product categories. This cross-category dependence is a direct consequence of the fact that the model implicitly defines a general utility function over all \( 2^N \) possible market baskets.

APPLICATION

In this section, we apply the multivariate logistic market basket model to the analysis of basket choice involving four paper goods categories. We show that the model predicts choice better than a simpler model that assumes independence in choice across the categories. This analysis shows that marketing mix actions that increase choice probability in one paper goods category impact choice probabilities in the remaining categories. However, the magnitude of these effects—measured from the perspective of cross-price elasticities—is small.
Data Description

The data are taken from a purchase panel of 170 households in the Toronto, Canada metropolitan area over a 2-year period. Purchases are recorded for four paper goods categories: paper towels, toilet paper, facial tissue, and paper napkins. These data were selected for analysis because the four categories contain national brand names and private labels that cut across product category boundaries (see Russell and Kamakura, 1997 for details). Moreover, paper goods products are typically bulky and are usually located in the same area of a grocery store. For these reasons, there is reason to suspect a priori that choice across these categories will be correlated.

The analysis was conducted by splitting the data into three consecutive periods. The first 30 weeks of the data were used to create household-specific category loyalty variables (defined subsequently). The remainder of the data was split into two sets: a model calibration period (2,578 baskets over 41 weeks) and a holdout period (822 baskets over a subsequent 30 weeks). Price levels and inter-purchase times are very similar across the two time periods (Table 2). However, average loyalty values differ across calibration and holdout time periods because 61 households have no records during the weeks covered by the holdout data.

The market basket distribution shows a highly skewed pattern. These data, listed in Table 3, identify each basket in terms of its contents: P = paper towels, T = toilet paper, F = facial tissue, and N = paper napkins. As might be expected, smaller baskets occur

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Household Data Summary</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Calibration Data</strong></td>
</tr>
<tr>
<td>Category loyalty</td>
</tr>
<tr>
<td>(log weeks)</td>
</tr>
<tr>
<td>Time since last purchase</td>
</tr>
<tr>
<td>(log weeks)</td>
</tr>
<tr>
<td>Category price index</td>
</tr>
<tr>
<td>(dollars per equivalent unit)</td>
</tr>
<tr>
<td><strong>Holdout Data</strong></td>
</tr>
<tr>
<td>Category loyalty</td>
</tr>
<tr>
<td>(log weeks)</td>
</tr>
<tr>
<td>Time since last purchase</td>
</tr>
<tr>
<td>(log weeks)</td>
</tr>
<tr>
<td>Category price index</td>
</tr>
<tr>
<td>(dollars per equivalent unit)</td>
</tr>
</tbody>
</table>

Note: Calibration data consists of 2,578 baskets of 169 households over 41 weeks period. Holdout data consists of 822 baskets of 108 households over a subsequent 30 week period. Differences in category loyalty across these data periods are due to the fact that 61 households have no data during the holdout weeks. Standard deviations are shown in parentheses. See text for variable definitions.
### Table 3

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>20.8</td>
<td>28.7</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>16.7</td>
<td>14.6</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>17.6</td>
<td>21.5</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5.7</td>
<td>3.9</td>
</tr>
<tr>
<td>PT</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>8.7</td>
<td>6.0</td>
</tr>
<tr>
<td>PF</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>7.5</td>
<td>11.0</td>
</tr>
<tr>
<td>FN</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1.7</td>
<td>1.1</td>
</tr>
<tr>
<td>TF</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>6.7</td>
<td>5.0</td>
</tr>
<tr>
<td>TN</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2.1</td>
<td>0.9</td>
</tr>
<tr>
<td>FN</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.4</td>
<td>0.9</td>
</tr>
<tr>
<td>PTF</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>6.2</td>
<td>4.3</td>
</tr>
<tr>
<td>PTN</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1.8</td>
<td>0.9</td>
</tr>
<tr>
<td>PFN</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>TFN</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1.2</td>
<td>0.2</td>
</tr>
<tr>
<td>PTFN</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1.3</td>
<td>0.7</td>
</tr>
<tr>
<td>Total Number of Baskets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2578</td>
<td>822</td>
</tr>
</tbody>
</table>

**Note:** P = paper towels, T = toilet paper, F = facial tissue, and N = paper napkins. Basket size is the total number of categories in the basket. The null basket (no categories purchased) is not included because null basket purchases were not recorded in the data collection process. Basket codes correspond to Figure 1.

...much more frequently than larger baskets. Note that Table 3 does not contain a frequency count for the null basket (i.e., no purchase in any of the four categories). This is due to the fact that the dataset was constructed conditional upon the household buying at least one of the four paper goods categories on a given shopping trip. In analyzing these data, we make an adjustment for the fact that the null basket is never observed.

Before the formal analysis is discussed, it is interesting to examine the conclusions that would be drawn from an affinity analysis (Brand and Gerristen, 1998). Affinity analysis compares the observed market basket distribution to a hypothetical distribution that would be observed if the presence of a category in a basket is statistically independent of the presence of any other category. By identifying categories that co-occur more (or less) frequently than expected, retail policy recommendations (such as store layout) are developed. Intuitively, affinity analysis can be regarded as a method of clustering categories into groups that are purchased on the same shopping occasion.

To illustrate affinity analysis for our data, we fit a main effects log-linear model to the Table 3 calibration basket counts and obtained a forecasted distribution. Because main effects log-linear models assume independence across the factors in a contingency table, the forecasts of this model are equivalent to a affinity analysis benchmark. This benchmark distribution, labeled "Independence" in Figure 1, is significantly different from the observed basket distribution as judged by a chi-squared test ($p < .0001$). Notice that
baskets of size two appear less often than expected, whereas baskets of size one generally occur more often than expected.

Accordingly, affinity analysis would argue that the four paper goods categories act as substitutes: the presence of one category in the basket decreases the likelihood that another category will be in the basket. As we show subsequently, these conclusions are not correct. The key problem is that affinity analysis is subject to biases because it ignores both consumer heterogeneity and marketing mix effects.

Model Specification

To specialize the market basket model to the paper goods dataset, we define the household characteristic \( HH_{kt} \) and marketing mix \( MIX_{kt} \) terms of Equation (6). We assume that characteristics of household \( k \) can be expressed as

\[
HH_{kt} = \delta_1 \log[TIME_{kt} + 1] + \delta_2 \text{LOYAL}_{kt}
\]

(9)

where \( TIME_{kt} \) is the time in weeks since the household's last category purchase and \( \text{LOYAL}_{kt} \) is a loyalty variable that adjusts for the household's long-run propensity to buy the category.\(^1\) We define \( \text{LOYAL}_{ik} = \log([n(i,k) + .5]/[n(k) + 1]) \) where \( n(i,k) \) is the number of product category \( i \) purchases across the household's \( n(k) \) purchase events in the
initial 30 weeks of the dataset. Because TIME_{ikt} is a surrogate for category inventory and LOYAL_{ijk} is a measure of interest in the product category, we expect both δ_{i} and δ_{j} to be positive.

In this analysis, the marketing mix of the basket model is defined as

\[ \text{MIX}_{ikt} = \gamma_{i} \log[\text{PRICE}_{ikt}] \]  

(10)

where PRICE_{ikt} is a price index for category i at time t. (The dependence on household k is due to the fact that different households face different marketing environments.) The index is a weighted average price taken across all stock keeping units (SKU’s) in the category. Weights are long-run volume shares for the SKU’s for the entire purchase panel over the first 30 weeks of the dataset. Price is defined in terms of dollars per equivalent unit. Category level promotional variables (feature and display) are excluded from the analysis due to high correlation with price. Consequently, the price coefficient in the model captures both regular price and promotional effects. We expect γ_{i} to be negative.

A key component of the model is the specification of the θ_{ijk} cross-effect terms. We model these effects as

\[ \theta_{ijk} = \delta_{ij} + \phi \text{SIZE}_{k} \]  

(11)

where the basket size loyalty variable SIZE_{k} is set to the mean number of categories per trip chosen by household k during the initial 30-week period of the data. To force the cross effects to be symmetrical within each household, we impose the constraint that the δ_{ij} be symmetrical with respect to categories i and j. Equation (11) accounts for the fact that shopping style will have an impact on the magnitude of the cross effects. If a household shops infrequently, it is more likely to buy a large basket of goods on each trip. Basket size loyalty measures this type of behavior. Clearly, households that have larger baskets on a typical shopping trip should exhibit larger cross effects in the market basket model. Accordingly, we expect ϕ to be positive.

**Model Calibration**

Because these data were constructed in such a way as to exclude the null basket from consideration, it is necessary to slightly alter the form of the basket model. Recall that Equations (5) and (6) assume that all 2^N possible baskets are available for selection. However, to analyze these data, we need the form of the market basket distribution, conditional upon the knowledge that purchases are made in at least one category. That is, we need to constrain the basket choice model to the 2^N - 1 alternative baskets that can be observed in these data.

Given the form of Equation (5), this constrained basket choice model is extremely easy to infer. Let 0 denote the null basket. Then, using (5), the probability that a basket contains at least one product category is
\[ \Pr(B(k,t) = 0) = \sum_{b=0}^{b^*} \exp[\mu(b^*,b,k,t)]/\sum_{b=0}^{b^*} \exp[\mu(b^*,b,k,t)] \]  

(12)

where the numerator runs over all baskets that are not empty. Accordingly, by taking the ratio of Equation (5) to Equation (12), we find that the probability of selecting basket b, given that b is not the null basket, is

\[ \Pr(B(k,t) = b \mid B(k,t) \neq 0) = \frac{\exp[\mu(b,k,t)]}{\sum_{b=0}^{b^*} \exp[\mu(b^*,k,t)]} \]  

(13)

where the denominator runs over all baskets that are not empty. In essence, all we need do is retain the form of the original basket model, but exclude the null basket from the possible alternatives. Notice that the parameters in (13) are the same as the parameters in the original model. Hence, we will obtain consistent estimates of all parameters in the full basket model, despite the fact that we do not observe the selection of the null basket.

Parameter estimation is straightforward, again due to the form of the basket model. Note that the formal structure of the market basket likelihood function is identical to a single category logit likelihood function defined over \(2^N-1 = 15\) possible alternatives. That is, given our basket model, we can approach model calibration as if the household made one selection out of 15 alternatives on each purchase occasion using the probabilities defined by Equations (13) and (6). Accordingly, model parameters were computed using a standard multinomial logit maximum likelihood estimation algorithm.\(^2\) The form in which explanatory variables enter the model follows the general structure of Equation (6).

**Model Comparison**

Before interpretation of the multivariate logistic basket model is attempted, two key questions should be addressed. First, does the multivariate logistic basket model represent an improvement in forecasting ability relative to an analysis of each category separately? Second, is there any evidence that the multivariate logistic model is correctly specified? Both questions are important because they deal with the managerial usefulness of the results. To address these questions, we compare the performance of various types of market basket models with respect to the paper goods data (Table 4).

Two classes of models are considered: Multivariate Logistic models, and Benchmark models. The Multivariate Logistic models are different versions of the basket model proposed in this research. Model C1 (Independence) assumes that all cross effects \(\theta_{ik}\) are zero. This model is equivalent to fitting a separate logit model for each of the four paper goods categories. Model C2 (Simple Cross) assumes that the cross effects \(\theta_{ik}\) do not vary across households. Model C3 (Full Cross Effects) allows cross effects to vary with respect to basket size loyalty, as specified by Equation (11).

The Benchmark models do not follow the logic of the multivariate logistic model discussed earlier. Instead, they use different specifications of category price variables to add cross category effects to the model. Models B1 and B2 begin with the Independence model (C1) and add price terms. In model B1, the MIX variable in the conditional logit
TABLE 4

Comparison of Alternative Basket Models

<table>
<thead>
<tr>
<th>Model Code</th>
<th>Model Description</th>
<th>Parm</th>
<th>Calibration Log Likelihood</th>
<th>Bayesian Information Criterion</th>
<th>Holdout Log Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate Logistic Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>Independence (no cross effects)</td>
<td>16</td>
<td>-5,170.64</td>
<td>10,466.96</td>
<td>-1,521.98</td>
</tr>
<tr>
<td>C2</td>
<td>Simple cross effects ($\delta_{ij}$ only)</td>
<td>22</td>
<td>-5,124.49</td>
<td>10,421.78</td>
<td>-1,505.23</td>
</tr>
<tr>
<td>C3</td>
<td>Full cross effects ($\phi_{ij} + \delta_{ij}$)</td>
<td>23</td>
<td>-5,052.61</td>
<td>10,285.88*</td>
<td>-1,498.74*</td>
</tr>
<tr>
<td>Benchmark Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>Model C1 + prices of all categories in main effects</td>
<td>28</td>
<td>-5,158.31</td>
<td>10,536.55</td>
<td>-1,528.44</td>
</tr>
<tr>
<td>B2</td>
<td>Model C1 + prices of all categories in cross effects</td>
<td>18</td>
<td>-5,176.71</td>
<td>10,389.42</td>
<td>-1,523.18</td>
</tr>
<tr>
<td>B3</td>
<td>Model C3 + prices of all categories in main effects</td>
<td>35</td>
<td>-5,042.54</td>
<td>10,360.00</td>
<td>-1,505.73</td>
</tr>
<tr>
<td>B4</td>
<td>Model C3 + prices of all categories in cross effects</td>
<td>29</td>
<td>-5,045.67</td>
<td>10,319.13</td>
<td>-1,502.31</td>
</tr>
</tbody>
</table>

Note: Parm is number of parameters. Asterisk denotes the best model according to the Bayesian Information Criterion (BIC) and Holdout Log Likelihood (HLL). The BIC is based upon the fit to the calibration data, while the HLL is based upon the fit to the holdout data. For the BIC, the best model has the smallest BIC value. For the HLL, the best model has the largest HLL value.

models [Equations (1) and (2)] is assumed to depend on the prices of all categories—not just the price of the given category. In model B2, Equation (11) is modified to

$$\theta_{ik} = \tau_{ij} \log[\text{PRICE}_{ik}] \log[\text{PRICE}_{jk}]$$

(14)

where $\tau_{ij}$ is symmetrical in categories i and j. Models B3 and B4 are constructed in an analogous way, but use the Full Cross Effects model (C3) as the base. Model B3 adds cross price terms to the conditional logit models [Equations (1) and (2)]. Model B4 modifies Equation (11) to

$$\theta_{ik} = \delta_{ij} + \phi_{ik} + \tau_{ij} \log[\text{PRICE}_{ik}] \log[\text{PRICE}_{jk}]$$

(15)

where $\delta_{ij}$ and $\tau_{ij}$ are symmetrical in categories i and j. In each case, Besag’s (1974) theorem is used to derive a corresponding market basket model (in logit form), which is then estimated. Because each of these models allows cross-category effects to be represented in a different way, they serve as reference points to the proposed multivariate logistic model specification developed earlier.

In Table 4, we use both the Bayesian Information Criterion (BIC) and the log likelihood in the holdout data (HLL) to select the best model for the paper goods data. The BIC adjusts the log likelihood in the calibration data for the number of parameters estimated, whereas the HLL uses model forecasts to construct the log likelihood of the
### Table 5

Parameter Estimates for Full Cross Effects Model

<table>
<thead>
<tr>
<th></th>
<th>Paper Towels</th>
<th>Toilet Tissue</th>
<th>Facial Tissue</th>
<th>Paper Napkins</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base Level Parameters</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>.4750&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.3100&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-.4115&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.2500</td>
</tr>
<tr>
<td></td>
<td>(.2734)</td>
<td>(.1683)</td>
<td>(.2139)</td>
<td>(.6424)</td>
</tr>
<tr>
<td>Category loyalty</td>
<td>1.834&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.667&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.540&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.849&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(.1305)</td>
<td>(.1108)</td>
<td>(.1086)</td>
<td>(.1383)</td>
</tr>
<tr>
<td>Time since last purchase</td>
<td>.3097&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.1837&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.1694&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.8843&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(.0701)</td>
<td>(.0712)</td>
<td>(.0672)</td>
<td>(.0946)</td>
</tr>
<tr>
<td>Category price index</td>
<td>-.7300&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-.5485&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-.9571&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-1.240&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(.3761)</td>
<td>(.2014)</td>
<td>(.4123)</td>
<td>(.7360)</td>
</tr>
<tr>
<td><strong>Cross Effect Parameters (δ&lt;sub&gt;i&lt;/sub&gt;)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basket size loyalty (ϕ)</td>
<td>.6421&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.6421&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.6421&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.6421&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(.0592)</td>
<td>(.0592)</td>
<td>(.0592)</td>
<td>(.0592)</td>
</tr>
<tr>
<td>Paper towels cross effects (δ&lt;sub&gt;ij&lt;/sub&gt;)</td>
<td>- .6568&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- .8635&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- 1.174&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- .7249&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(.1807)</td>
<td>(.1800)</td>
<td>(.1970)</td>
<td>(.1961)</td>
</tr>
<tr>
<td>Toilet tissue cross effects (δ&lt;sub&gt;ij&lt;/sub&gt;)</td>
<td>- .6568&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- .7633&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- 1.373&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- 1.373&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(.1807)</td>
<td>(.1787)</td>
<td>(.1990)</td>
<td>(.1990)</td>
</tr>
<tr>
<td>Facial tissue cross effects (δ&lt;sub&gt;ij&lt;/sub&gt;)</td>
<td>- .8635&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- .7633&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- 1.373&lt;sup&gt;a&lt;/sup&gt;</td>
<td>- 1.373&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(.1800)</td>
<td>(.1787)</td>
<td>(.1990)</td>
<td>(.1990)</td>
</tr>
</tbody>
</table>

Note: The basket size loyalty coefficient is not category specific. Only one coefficient is estimated for the model. The δ<sub>i</sub> cross effect parameters are constrained to be symmetrical. Standard errors of parameters are shown in parentheses. Statistical significance is denoted as <sup>a</sup>(.05 level or better) and as <sup>b</sup>(.10 level or better).

Holdout data. Both criteria are commonly used in marketing science to select among competing models. The smallest BIC value and the largest HLL value identify the best model.

The clear conclusion is that the Full Cross Effects model (C3) best represents the choice process. Because the Full Cross Effects model is better than the Independence model (no cross effects), we can conclude cross-category information is important. That is, choices across categories are correlated. Moreover, the superiority of the Full Cross Effects model over all the benchmark models provides strong evidence that the multivariate logistic approach discussed earlier is a reasonable way of capturing these cross-category effects. For this reason, the remainder of the discussion is focused on the Full Cross Effects model.

### Parameter Estimates

The parameter estimates for the Full Cross Effects model are presented in Table 5. Setting aside the category-specific intercepts, all parameters are statistically significant and have the expected signs: negative for all price coefficients, and positive for coeffi-
Table 6

<table>
<thead>
<tr>
<th></th>
<th>Paper Towels</th>
<th>Toilet Tissue</th>
<th>Facial Tissue</th>
<th>Paper Napkins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper towels</td>
<td>—</td>
<td>.3192(^a)</td>
<td>.1125</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(.1205)</td>
<td>(.1200)</td>
<td>(.1408)</td>
<td></td>
</tr>
<tr>
<td>Toilet tissue</td>
<td>.3192(^a)</td>
<td>—</td>
<td>.2128(^b)</td>
<td>.2511(^b)</td>
</tr>
<tr>
<td></td>
<td>(.1205)</td>
<td>(.1183)</td>
<td>(.1413)</td>
<td></td>
</tr>
<tr>
<td>Facial tissue</td>
<td>.1125</td>
<td>.2128(^b)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(.1200)</td>
<td>(.1183)</td>
<td>(.1442)</td>
<td></td>
</tr>
<tr>
<td>Paper napkins</td>
<td>— .1975</td>
<td>.2511(^b)</td>
<td>— .3972(^a)</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(.1408)</td>
<td>(.1413)</td>
<td>(.1442)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Values shown are \(\theta_{ijk} = \delta_j + \delta_k \text{SIZE}_{ik}\) where \text{SIZE}_{ik} is set to the mean number of categories per trip across all households (\text{SIZE}_{ik} = 1.54). The standard errors shown in parentheses are inferred from the results reported in Table 5. Statistical significance is denoted as \(^a\)(.05 level or better) and as \(^b\)(.10 level or better).

...coefficients corresponding to loyalty, time since last purchase, and basket size loyalty. It should be noted that the parameters in each column of the table correspond to one of the four conditional logit models defined by Equations (3) and (4). However, the parameters collectively define the multivariate logistic basket model of Equation (6). Note that we are able to obtain estimates of all parameters in the model, despite the lack of information on null baskets.

Interpretation of the demand relationships depicted by the cross effects is difficult because basket size loyalty varies across households. To gain insight into the cross-category effects, we present the cross effects for a typical household in Table 6. These values were computed by replacing \text{SIZE}_{ik} in Equation (11) by the value 1.54, the mean of \text{SIZE}_{ik} across all households. Three of the categories (paper towels, toilet tissue, and facial tissue) have positive cross effects and consequently act as demand complements. The clear outlier is paper napkins, which has a mixture of positive and negative relationships with respect to the remaining categories. This pattern is consistent with the fact that the paper napkin category has the longest interpurchase time of the four paper goods categories in this study (Table 2).

In reading Table 6, it is important to bear in mind that the magnitude of these effects differs across households. Because the coefficient on \text{SIZE}_{ik} is positive, households that tend to buy more categories per shopping trip on average will have larger coefficients and are more likely to exhibit complementarity in cross-category relationships. In contrast, households that buy fewer categories per shopping trip on average will have smaller coefficients and are more likely to exhibit substitution across categories. As we show subsequently, these same effects are evident when cross-category price elasticities are calculated.

Cross-Category Price Elasticities

From a managerial perspective, the most interesting aspects of this research are cross-category price elasticities. In Table 7, we display the percentage change in the
Table 7

<table>
<thead>
<tr>
<th></th>
<th>Paper Towels</th>
<th>Toilet Tissue</th>
<th>Facial Tissue</th>
<th>Paper Napkins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper towels</td>
<td>-.416</td>
<td>-.021</td>
<td>-.018</td>
<td>.005</td>
</tr>
<tr>
<td>Toilet tissue</td>
<td>-.031</td>
<td>-.308</td>
<td>-.027</td>
<td>-.019</td>
</tr>
<tr>
<td>Facial tissue</td>
<td>-.016</td>
<td>-.017</td>
<td>-.581</td>
<td>.018</td>
</tr>
<tr>
<td>Paper napkins</td>
<td>.009</td>
<td>-.024</td>
<td>.036</td>
<td>-.939</td>
</tr>
</tbody>
</table>

Note: Matrix displays the percentage change in the choice share of the row category with respect to a one percent increase in the price of the column category. Values in the table were computed using the parameters of Table 5 and the aggregate elasticity expressions in Appendix B.

choice share of the row category with respect to a one percent change in the price of the column category. These elasticities are computed using the forecasts of the Full Cross Effects model and an elasticity formula developed in Appendix B. The elasticities take into account consumer heterogeneity and are computed with respect to the long-run choice shares of the entire market. Consequently, they can be interpreted as the pattern of cross-category price elasticities that a retailer would observe in a typical week.

Several aspects of Table 7 are noteworthy. First, as might be expected, own-price effects (the diagonal of the matrix) are less than one in absolute value—implying inelastic demand for the four paper goods categories. (In contrast, single-category studies of brand price competition typically show elastic demand.) Second, most of the cross elasticities are negative—implying complementarity. Again, the exception is the paper napkins category, which acts as a substitute with respect to paper towels and facial tissue. These cross elasticities are asymmetric (despite the symmetry imposed upon the cross-category coefficients $\theta_{jk}$). In general, the properties of this elasticity matrix are reasonable given that our analysis examines choice across product categories.

The most striking aspect of this analysis is the small size of the cross-price effects. Although patterns of complementarity and substitution are present in Table 7, cross-category spillover effects due to price are not very important in terms of market-level category choice shares. However, recall that the Full Cross Effects model (used in computing the cross-category price elasticities) fits the data better than an Independence model, which assumes that no cross-category correlation in choice exists. Moreover, this improvement can be detected when forecasting to a holdout data period (Table 4). Taken together, these findings suggest that cross-category correlation in choice exists, but is due primarily to the consumer’s shopping style. Consumers do buy paper goods categories together on the same shopping trip, but variation of price in one category has a modest impact upon sales in other paper goods categories.4

Research Implications

It is important to place these conclusions in the proper context. The key attributes linking the four paper goods categories are proximity in store layout and cross-category
TABLE 8

Impact of Basket Size on Cross-Price Elasticity

<table>
<thead>
<tr>
<th></th>
<th>Small Basket Consumers</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Paper Towels</td>
<td>Toilet Tissue</td>
<td>Facial Tissue</td>
<td>Paper Napkins</td>
</tr>
<tr>
<td>Paper Towels</td>
<td>-.505</td>
<td>-.008</td>
<td>.007</td>
<td>.021</td>
</tr>
<tr>
<td>Toilet Tissue</td>
<td>-.012</td>
<td>-.386</td>
<td>.004</td>
<td>-.004</td>
</tr>
<tr>
<td>Facial Tissue</td>
<td>.006</td>
<td>-.002</td>
<td>-.683</td>
<td>.034</td>
</tr>
<tr>
<td>Paper Napkins</td>
<td>.036</td>
<td>-.005</td>
<td>.068</td>
<td>-1.034</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Large Basket Consumers</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Paper Towels</td>
<td>Toilet Tissue</td>
<td>Facial Tissue</td>
<td>Paper Napkins</td>
</tr>
<tr>
<td>Paper Towels</td>
<td>-.374</td>
<td>-.028</td>
<td>-.030</td>
<td>-.002</td>
</tr>
<tr>
<td>Toilet Tissue</td>
<td>-.038</td>
<td>-.275</td>
<td>-.037</td>
<td>-.026</td>
</tr>
<tr>
<td>Facial Tissue</td>
<td>-.026</td>
<td>-.023</td>
<td>-.531</td>
<td>.010</td>
</tr>
<tr>
<td>Paper Napkins</td>
<td>-.004</td>
<td>-.033</td>
<td>.020</td>
<td>-.893</td>
</tr>
</tbody>
</table>

Note: Matrices display the percentage change in the aggregate choice share of the row category with respect to one percent increase in the price of the column category. The median household buys 1.45 paper goods categories per trip. Households below the median are classified as “Small Basket Consumers.” Households above the median are classified as “Large Basket Consumers.”

brand names. These features are apt to be much weaker determinants of demand interdependence than true consumption complementarity. Earlier work by Bodapati and Srinivasan (1999), using a set of categories with no obvious consumption complementarities, also found weak cross-category demand effects. Indeed, the only market basket study to date that has found relatively large cross-category elasticities examined strong consumption complements such as cake mix and frosting (Manchanda et al., 1999). An emerging generalization may be that strong consumer perceptions of category relatedness are necessary before cross-category price spillover effects will be observed in a market basket context.

It is important to understand that market basket models are designed to forecast choice behavior within the grocery store. In effect, the consumer is assumed to be already in the store and variables such as category price level are used to predict which basket of categories will be selected. Because market basket models do not address store traffic effects, the lack of strong cross-category price elasticities in this study should not be interpreted as evidence that category pricing activity has no impact on the types of baskets that the retailer will observe in particular week. It is entirely possible that the major impact of category pricing is on store choice—an aspect of choice behavior that is not modeled in this research. Because there is strong evidence that preferences are correlated across product categories (Russell and Kamakura, 1997), cross-category demand effects may be largely determined by week-to-week fluctuations in the set of households buying in a particular store.

In fact, our data do provide an indirect indication that prices could influence store choice. In Table 8, we repeat the elasticity analysis of Table 7, but separate households into two groups, depending upon how many paper goods categories are purchased on a typical shopping trip. As might be expected, the small basket group shows more substi-
tution across categories, than the large basket group. Again, these cross effects are small. The key finding is that own price effects are uniformly larger for the small basket group than for the large basket group. (Compare the diagonal terms of the two elasticity matrices in Table 8.) It may in fact be the case that paper goods promotions will differentially attract consumers who tend to buy smaller market baskets. This type of “cherry picking” behavior is not advantageous to the retailer, but may be a stable characteristic of consumer behavior. Bell and Lattin (1998) also provide evidence that small basket consumers exhibit higher price sensitivity with respect to category choice than large basket consumers.

CONCLUSIONS

This research develops a new approach to market basket construction based upon the notion that choice in one category impacts choices in other categories. The approach assumes that the researcher can specify the probability that a consumer chooses one category in the basket, given information on the actual choice outcomes in all other categories. We show that by using these conditional choice models, it is possible to infer the market basket distribution that explains purchasing in all categories. We applied the approach to the analysis of choice in four paper goods categories. Substantively, we showed that choice across four paper goods categories is correlated, but that the within-store magnitudes of cross-category price effects are modest.

Methodological Contribution

The model developed here has a number of advantages. Although the logic behind the model is consistent with the sequential choice approach of Harlam and Lodish (1995), it does not require the researcher to actually observe the order in which choices are made. Given the fact that decision sequences are rarely recorded in consumer purchase histories, this feature gives the proposed approach much greater applicability. Moreover, the form of the model developed here is computationally very attractive. Because the multivariate logistic distribution shares certain similarities with logit choice models, estimation algorithms developed for single category logit analysis can be readily adapted to the analysis of choice in multiple product categories. In effect, by thinking of the process as one of choosing baskets rather than individual categories, we are able to recast a multiple category decision model into a single category choice framework. The result is a analytically-tractable market basket choice model that allows general patterns of choice dependence (complementarity, independence, or substitution).

Managerial Contribution

Despite the failure to detect strong cross price effects across the four paper goods categories, this model may nevertheless prove useful in studying the impact of cross-
category marketing activity for other groups of categories. If the proposed market basket analysis were conducted with a large number of categories, it would be possible to describe the retail category assortment in terms of the strength of cross-category relationships. Blocks of categories with strong complementarity relationships are particularly interesting because promotional activity in one category can simultaneously increase sales in other categories within the same block. If the retailer were to discover only weak cross-category effects, then market baskets can be forecasted using independent category-level choice models. Under this scenario, understanding the long-run preferences of consumers and predicting store choice among these consumers are more important in forecasting than modeling cross-category choice correlations at the time of purchase.

It is likely that cross-category price effects are strongly affected by perceived relatedness of product categories. Our inability to find strong cross-category effects for the paper goods categories provides evidence that cross-category branding and physical proximity in the store do not provide point-of-purchase cues which stimulate cross-category purchasing. Rather, true consumption complementarity (such as the joint usage of detergent and fabric softener) appears to be required to generate strong cross-category effects. However, retailers potentially can enhance perceptions of cross-category relatedness by using merchandising tools (point-of-purchase materials, cross-category coupons, and creative store layout) to suggest cross-category consumption goals to the consumer. The methodology developed in this research could be a useful tool in measuring the success of such retailer actions in building larger market baskets. Given the strong interest by retailers in category management, the development and assessment of marketing policies that promise inter-category synergies is clearly of interest.

**Future Work**

There are a number of limitations to the current model, all of which provide opportunities for future research. As noted earlier, the model developed here ignores store choice, assuming that the consumer is already in the store and ready to make purchases. The model also does not identify the particular product chosen when a category is selected nor provide an estimate of purchase volume. Both these issues can be addressed in the conditional choice framework, but would require the development of a nested probability structure using different types of conditional choice distributions. These extensions are potentially very important because they provide alternative perspectives on the market basket choice decision.

Finally, as the number of categories becomes large, the approach taken in our research will clearly become infeasible. A typical retailer in the United States carries approximately 31,000 items, divided into 600 product categories (Kahn and McAlister, 1997). Any attempt to build a choice model that explicitly enumerates the $2^{600} = 10^{181}$ possible baskets will obviously fail. However, another route is open to the researcher. It is possible to estimate each of the full conditional models individually (with side conditions to ensure mutual consistency) and then to use Markov Chain Monte Carlo simulation methodologies to forecast the full market basket distribution (see, e.g., Gilks et al., 1996; Gelman et al.,
1996). This general procedure (which effectively reduces a $2^N$-sized problem to an $N$-sized problem) may allow the development of a practical forecasting tool for large market baskets.

Acknowledgment: The authors thank Professor Andrew Mitchell, Director of the Canadian Centre for Marketing Information Technologies, for providing access to the data used in this research. The authors also thank Greg Allenby, V. Srinivasan, Osnat Stamer and Doyle Weiss for many helpful comments. This research was supported by the College of Business Summer Grant Program and by a special grant from Mr. Robert Jensen to the College of Business.

NOTES

1. To emphasize the simplicity of model calibration for the applied researcher, we capture all consumer heterogeneity using observed variables for category loyalty and interpurchase time. More sophisticated statistical approaches—such as latent class models (Kamakura and Russell, 1987) and random coefficient models (Jain, Vilcassim and Chintagunta, 1994)—would be required to represent unobserved consumer heterogeneity.

2. The tractability of the multivariate logistic distribution is a key feature of the methodology developed here. In contrast, a market basket model based upon the multivariate probit (Manchanda et al., 1999) requires the use of Markov Chain Monte Carlo simulation techniques to evaluate the high dimensional integral defining purchase probability.

3. The superiority of the Full Cross Effects model over the Independence model can also be shown using the classical likelihood ratio test.

4. The importance of the cross-category effects can be measured using the $\rho^2$ statistic. This statistic is defined as $(LL(base)-LL(m))/LL(base)$ where LL(m) is the log likelihood of the model under consideration and LL(base) is the log likelihood of a base model which assumes that all baskets are equally likely. The value of $\rho^2$ runs between zero and one, with higher values indicating better fit. Using information on fit in the holdout data sample, we find the following values: Independence = .316, Simple Cross Effects = .324 and Full Cross Effects = .327. Evidently, cross-category effects are present in paper goods basket choice behavior, but the incremental improvement in fit due to these effects is small.

Appendix A:

Relationship Between Full Conditional and Joint Distributions

The factorization theorem of Besag (1974) allows the researcher to verify consistency of the full conditionals and to derive the form of the implied joint distribution. Let $X = \{x(1), x(2), \ldots, x(N)\}$ be any basket of category choices. Let $f(X)$ denote the joint distribution of the random variables $x(1), x(2), \ldots, x(N)$. (This can be interpreted as the probability of observing a basket with contents $X$.) Define the vector $0 = \{0, 0, \ldots, 0\}$ as the null basket and let $f(0)$ be the probability associated with the null basket. For any
Cross Category Dependence

permutation of the category labels, the joint (market basket) distribution is implicitly defined by

\[ f(X)/f(0) = k(1)^*k(2)^* \ldots *k(N) \tag{A1} \]

where \( k(i) = g^i(x(i))/g^i(0) \) depends upon the full conditional distributions

\[ g^i(\cdot) = f_{-i}(x(1), \ldots, x(i-1), 0, \ldots, 0). \tag{A2} \]

An explicit expression for \( f(X) \) can be worked out using the fact that the summation of (A1) over all market baskets \( X \) is equal to \( 1/f(0) \).

Because the order in which the categories are arranged in \( X \) is arbitrary, the full distribution will not be unique unless all permutations of category labels yield the same joint distribution according to Equations (A1) and (A2). As Besag (1974) notes, this generally requires the researcher to place restrictions on the form of the full conditional distributions. In the present application, symmetry of the cross effects \( \theta_{ij} \) ensures that the form of the joint distribution is invariant with respect to label permutations.

Appendix B:
Derivation of Own and Cross-Category Price Elasticities

Expressions for price elasticities depend upon the specification of the market basket model. Following the form of the cross-effects model discussed in the text, we assume that the probability that consumer \( k \) buys basket \( b \) at time \( t \) is given by

\[ \Pr(B(k,t) = b) = \exp{\mu(b,k,t)}/\Sigma_{b'} \exp{\mu(b'^*,k,t)} \tag{B1} \]

where

\[ \mu(b,k,t) = \Sigma_i \beta_i(i,b) + \Sigma_i HH_{ai} X(i,b) + \Sigma_i MIX_{ai} X(i,b) + \Sigma_{i<j} \theta_{jk} X(i,b) X(j,b) \tag{B2} \]

is the implied utility of a basket with contents \( \{X(1,b), X(2,b), \ldots, X(N,b)\} \). Here \( X(i,b) \) is a binary 0–1 variable that takes the value 1 when category \( i \) is present in basket \( b \). We assume that price enters into the utility expression (B2) only as \( MIX_{ikt} = \gamma_i \log[PRICE_{kbt}] \). All other terms in (B2)—including the symmetrical cross effects \( \theta_{ijk} = \phi[SIZE_{ai}] + \delta_{ij} \)—do not depend on price.
Consumer-Level Price Elasticities

We consider summations of $\exp\{\mu(b,k,t)\}$ over various subsets of market baskets. The notation $SB(\text{all})_{kt} = \sum_{b} \exp\{\mu(b^*,k,t)\}$ denotes a summation over all possible baskets (including the null basket). $SB(i)_{kt}$ denotes the summation of $\exp\{\mu(b,k,t)\}$ over all baskets $b$ containing category $i$. In addition, $SB(i,j)_{kt}$ denotes the summation of $\exp\{\mu(b,k,t)\}$ over all baskets $b$ containing both category $i$ and category $j$.

It is important to understand that price elasticities are defined relative to product categories, not with respect to market baskets. Define the probability of a consumer buying category $i$ on a shopping trip as

$$\Delta(i)_{kt} = SB(i)_{kt}/SB(\text{all})_{kt} \tag{B3}$$

Formally, this is the probability that the consumer chooses a basket that contains category $i$, regardless of which additional categories are present. Analogously, the probability that the consumer chooses a basket containing both category $i$ and category $j$ is

$$\Delta(i,j)_{kt} = SB(i,j)_{kt}/SB(\text{all})_{kt} \tag{B4}$$

In words, this is the probability that the selected basket contains both $i$ and $j$, regardless of which additional categories are present. These expressions emphasize the fact that the basket model can be thought of as a probability distribution over product category choices—not just a probability distribution over market baskets.

Using these expressions, we define the cross-category price elasticity $E(i,j)_{kt}$ as the percentage change in the probability of selecting category $i$ with respect to a one percent change in the price of category $j$. That is, $E(i,j)_{kt} = \partial(\log \Delta(i)_{kt})/\partial(\log \text{PRICE}_{jkt})$. Using this definition, Equations (B1) through (B4) imply that

$$E(i,i)_{kt} = \gamma_i (1 - \Delta(i)_{kt}) \tag{B5}$$

$$E(i,j)_{kt} = \gamma_i \Delta(j)_{kt} / S(i,j)_{kt} - 1, \quad i \neq j \tag{B6}$$

where $S(i,j)_{kt} = \Delta(i,j)_{kt}/\Delta(i)_{kt}\Delta(j)_{kt}$ is a measure of the association across the two product categories. In these expressions, we expect $\gamma_i$ and $\gamma_j$ to be negative. Accordingly, own elasticities are always negative. Cross elasticities can be positive or negative, depending upon the value of $[S(i,j)_{kt} - 1]$.

Aggregate Elasticities

To obtain aggregate price elasticities, we define the aggregate choice share of category $i$ as the mean $\Delta(i) = \sum_k \Delta(i)_k/N$ where $N$ is the total number of households. This expression is the expected choice share for category $i$ in week $t$ across the entire market.
We also define the aggregate cross-category price elasticity as $E(i,j)_t = \frac{\partial \log \Delta(i)_t}{\partial \log \text{PRICE}_{jt}}$ where \( \text{PRICE}_{jt} \) is interpreted as the price of category \( j \) in week \( t \) for each consumer in turn. Using Equations (B5) and (B6), we can derive the aggregate cross-price elasticities

\[
E(i,j)_t = \gamma[\sum_k \Delta(i)_t(1 - \Delta(i)_t)]/[\sum_k \Delta(i)_t]
\]  

(B7)

\[
E(i,j)_t = \gamma[\sum_k \{\Delta(i,j)_t - \Delta(i)_t \Delta(j)_t\}]/[\sum_k \Delta(i)_t], \ i \neq j
\]

(B8)

for a particular time point \( t \). The differences between these expressions and the individual price elasticities in (B5) and (B6) are due to aggregation over heterogeneous consumers.

An alternative procedure is to define elasticities with respect to the overall choice shares \( \Delta(i) = [\sum_k \Delta(i)_t]/N^* \), where \( N^* \) is the number of choice sets found in a particular data set. (\( N^* \) equals \( N \) times the average number of choices occasions per household.) These elasticities can be regarded as the typical aggregate elasticities that would be found in the market during a randomly selected week. Using this definition and Equation (B5) and (B6), we obtain the aggregate cross-price elasticities

\[
E(i,j) = \gamma[\sum_k \sum_t \Delta(i)_t(1 - \Delta(i)_t)]/[\sum_k \sum_t \Delta(i)_t]
\]

(B9)

\[
E(i,j) = \gamma[\sum_k \sum_t \{\Delta(i,j)_t - \Delta(i)_t \Delta(j)_t\}]/[\sum_k \sum_t \Delta(i)_t], \ i \neq j
\]

(B10)

In the text, we report these aggregate elasticities in the discussion of cross-category price effects. It should be noted that if all cross-effect parameters \( \theta_{ijk} \) in (B2) are zero, then all cross-price elasticities in (B10) will also be zero.

The aggregation procedures adopted here are similar in spirit to procedures advocated by Russell and Kamakura (1994) and Bucklin, Russell, and Srinivasan (1998). However, the expressions shown here are specific to the market basket model in (B1) and (B2).

**REFERENCES**


