A Cluster Analytic Approach to Market Response Functions

DONALD E. SEXTON, JR.*

INTRODUCTION

The estimation of marketing policy effects with time series data is a fairly well-mined area of marketing discourse. Use of such data bases to examine the differences in effects on various market segments has received relatively less attention (except [3]). This is surprising given the premise of market segmentation, namely that certain groups of buyers behave differently than others. If a market consists of segments, ideally one would think that the effects of marketing policies should be estimated segment by segment. Then the market response function would be simply the aggregate of the functions estimated separately for each segment. That is, let:

$$S_i = f_i(A, P, D)$$

where: S_i = brand sales in segment i,
A = measure of advertising level in market,
P = measure of price level in market, and
D = measure of deal level in market.

Then,

$$S = \sum_i S_i = \sum_i f_i(A, P, D)$$

where: S = brand sales in entire market.

Under a linearity assumption, the segment-by-segment approach yields the same regression coefficients as if the market response curve had been estimated by performing regressions on observations corresponding to the entire market. That is, consider a market composed of k segments and assume linear effects of advertising, price, and deals.

$$S_i = [A, P, D]\beta_i + \epsilon, \qquad i = 1, \cdots, k$$

where: β_i = vector of coefficients of marketing policy variables for segment i, and
ϵ = random disturbance.

The ordinary least squares estimates of the coefficients in the market response curve can easily be shown to be simply the sum of the coefficient estimates from the segment response curves when, as assumed to this point, the advertising, price, and deal magnitudes are identical for each segment. (Such an assumption holds to the extent shoppers in all segments are exposed to the same media mix and frequent the same mix of stores.) That is,

$$b = \sum b_i$$

where: b = vector of estimated coefficients for market, and
b_i = vector of estimated coefficients for segment i.

Therefore, if a linear sales response curve is assumed, estimation does not suffer if market policy effect differences among segments are ignored.

In the last five years, however, much research has suggested that *nonlinear* response functions may be more appropriate than linear functions [7]. Nonlinear functions, however, do not have the reproductive property. A sum of loglinear functions, for example, is not a loglinear function. Moreover, most research has concentrated on *brand share* models. If brand share rather than sales is employed as the dependent variable, even linear functions do not have the reproductive property. In such cases the researcher must make a choice—either to assume one response function for the entire market or to specifically tailor a response function for each segment and then aggregate the results. If market segments are identifiable, of appreciable size, and accessible, the segment-by-segment approach would appear to be more appropriate.

Another argument for a segment-by-segment approach concerns the independent variables in marketing response functions. The marketing mix faced by the consumers in one segment may differ from that faced by those in another segment. For example, the stores generally patronized by one segment may be different from those patronized by another segment, and so each segment may encounter different price and deal distributions. In such circumstances, any function fitted

* Donald E. Sexton, Jr. is Associate Professor of Business, Graduate School of Business, Columbia University. This work was performed under a research grant from the Graduate School of Business, Columbia University. The author wishes to thank Professor Neil Beckwith, Columbia University, for his helpful suggestions and comments (although any remaining errors are mine).

Journal of Marketing Research, Vol. XI (February 1974), 109–14

across all consumers will yield less efficient predictions than an approach consisting of estimating separate market response curves for each segment. This study is an attempt to obtain more accurate estimates of both market policy effects and sales by (1) disaggregating data into segments, (2) analyzing each segment separately, and then (3) aggregating the results.

THE DATA

The data base consists of all the purchases of all brands of regular coffee, instant coffee, and tea by a stationary sample of 569 families over a period of 2½ years (beginning in October, 1963). These panel data were obtained from the Family Survey Bureau of the *Chicago Tribune* and included brand purchased, price, amount by weight, store where purchased, and whether or not purchased under a deal. The usual caveats regarding panel data apply—in particular, the sample families may be somewhat more price-conscious than the general population so that the absolute effects of price estimated in this study can be extrapolated only with care. This reservation, however, does not limit the conclusions of this investigation which are concerned with the *relative* effects of marketing policies in various market segments.

Brand advertising expenditures on network television, spot television, and magazines and brand advertising lineage in local newspapers were obtained from several marketing research services: Leading National Advertisers, Broadcast Advertising Reports, Publishers Information Bureau, and Media Records. Network radio outlays were found to be negligible and so are not included in this study. The 130 weeks of data were arbitrarily divided into three intervals:

Weeks	*Purpose*
1–26	Classification of families
27–106	Main sample for parameter estimation
107–130	Validation sample for parameter estimation

CONSTRUCTION OF SEGMENTS

The main difficulty in pursuing a segment-by-segment estimated approach is the identification of the segments. Ideally one would like to define segments composed of consumers who are homogeneous with respect to their elasticities for the various marketing mix variables, i.e., consumers with the same functional form and parameter values for their individual marketing response functions. Since such information would generally be expensive and difficult to obtain for each consumer, a practical alternative is to employ commonly available variables that may be expected a priori to be related to their elasticities.

In this study, purchase habits were used to segment the panel families. A priori it was felt that such characteristics might indicate families who would react differently to advertising, price, and deals. Heavy users might be expected to be more sensitive to price and dealing and brand disloyal buyers to be more sensitive to advertising.

Table 1

RELATIVE CHARACTERISTICS OF CLUSTERS

Clustering method	*Cluster*	*Number of families*	*Usage rate*	*Brand 1 proportion of purchases*	*Private brand proportion of purchases*	*Chain proportion of store visits*
Naive	1	259	low	low		
	2	46	low	high		
	3	199	high	low		
	4	65	high	high		
Two-variable	1	437	low	low		
	2	40	low	high		
	3	65	high	low		
	4	27	high	high		
Four-variable	1	237	low	low	low	low
	2	155	high	medium	medium	high
	3	104	high	low	high	high
	4	73	high	high	low	low

Thirteen purchase habit variables were considered, based on the purchases of the product of interest (regular coffee) and its substitutes (instant coffee and tea) as well as the types of stores patronized:

Variable	*Description*
1	Pounds of regular coffee purchased.
2	Pounds of instant coffee purchased.
3	Pounds of tea purchased.
4	Brand 1 purchases as proportion of all regular coffee purchases.
5	Brand 2 purchases as proportion of all regular coffee purchases.
6	Brand 3 purchases as proportion of all regular coffee purchases.
7	Brand 4 purchases as proportion of all regular coffee purchases.
8	Brand 5 purchases as proportion of all regular coffee purchases.
9	Brand 6 purchases as proportion of all regular coffee purchases.
10	Private brand purchases as proportion of all regular coffee purchases.
11	Proportion of store visits to chain stores.
12	Proportion of store visits to affiliated independent stores.
13	Proportion of store visits to unaffiliated independent stores (generally Mom 'n Pop stores).

The first 26 weeks of data were used to calculate the values of these variables for each family.[1] These observations were excluded from the later estimation work.

[1] The usage rate variables (1, 2, and 3) were converted to proportions by dividing each by the largest amount found across all families. This procedure removed a large portion of the scale effect that would have been present in the cluster analyses.

Brand 1 (with the second largest market share) was selected as the brand for which sales would be predicted. Several approaches for dividing the sample into segments, each involving various sets of the purchase habit variables, were examined. The first method was naive: families were classified as loyal to brand 1 if their purchases of brand 1 as a proportion of all their coffee purchases was greater than .10, brand 1's average market share during the first 26 weeks. (This definition of brand loyalty may seem arbitrary, but it agrees with a finding reported by Frank and Green [2] that coffee consumers in a given brand loyalty group generally bought only one brand at a rate greater than the brand's overall market share.) A family was considered a heavy user if during these 26 weeks they bought more than 10 pounds, the median usage rate over all families.

The other segmentation procedures employed cluster analysis. In these approaches, the Mahalanobis-D^2 statistic was used as a measure of distance between families. The families were classified into 3, 4, 5, and 6 groups on the basis of sets of 2, 4, 5, and 13 of the purchase habit variables. These sets consisted, respectively, of variables 1 and 4; 1, 4, 10, and 11; 1, 2, 4, 10, and 11; and 1 through 13. The two and four variable sets seemed to uncover the most coherent cluster structures, i.e., clusters that appeared easiest to describe. When five or six clusters were specified, the group structures tended to consist of one large cluster and several smaller ones. For that reason and for compatibility with the naive approach, the four cluster structures defined by the two and four variable sets were selected for the estimation work described in this article.

There was one major difficulty in these cluster analyses. There was no clustering routine available that was capable of handling more than 200 objects. Therefore, for each of the several clustering approaches, three analyses had to be made, each dealing with every third family (to avoid any bias caused by date of entry to the panel). Then groups of three similar clusters—one from each of the three analyses—were aggregated to form the clusters discussed in the article. As a result, the clusters in this study are not as homogeneous as those that would have been produced by a simultaneous analysis of all 569 families. A larger clustering program might have produced clusters with better-fitting market response functions than those reported here.

Table 1 summarizes the results of these segmentation attempts. For the naive and two-variable approaches, the clusters can be generally described as: (1) low usage and low loyalty, (2) low usage and high loyalty, (3) high usage and low usage, and (4) high usage and high loyalty. For the four-variable approach, clusters 1 and 4 can be similarly named. Clusters 2 and 3, however, both consist of high usage families who are distinguished from other families by shopping more at chains and who are distinguished from each other by the proportions of their purchases represented by brand 1 and by private labels. In particular, cluster 3 families appear as if they may be more sensitive to prices than cluster 2 families because of their higher proportion of private brand purchases.

Table 2

F-VALUES FOR ALTERNATE MODELS

Cluster method	*Cluster*	*Linear*	*Loglinear*	*Negative–exponential*
All families		3.39[a]	2.41[b]	3.41[a]
Naive	1	1.61	.44	1.65
	2	.55	.42	.62
	3	1.87	.52	1.89
	4	3.75[a]	2.92[a]	3.77[a]
Two-variable	1	2.21[b]	1.46	2.26[b]
	2	1.21	.46	1.21
	3	8.03[a]	3.12[a]	7.90[a]
	4	3.11[a]	2.54[b]	2.45[b]
Four-variable	1	1.64	.45	1.67
	2	2.68[a]	.85	2.60[a]
	3	3.08[a]	1.01	3.12[a]
	4	4.66[a]	6.50[a]	4.31[a]

[a] Significant at .01 level.
[b] Significant at .05 level.

In preparation for the estimation work, the purchase records for the 80-week interval comprising the main sample and for the 24-week interval comprising the validation sample were aggregated by cluster. That is, purchases for all families in a cluster during a given week were summed to yield sales in that segment for that week.

BRAND SHARE MODELS

In previously reported work on this same data base [5, 6], a variety of market response functions were examined by performing ordinary least squares regressions on variables based on all the families together. Dependent variables considered included both sales and brand share. The independent variables explored included time, the dependent variable lagged one period, and advertising, pricing, and dealing all lagged up to two or three periods. In addition, three different forms of the relationship between these various combinations of variables were investigated: linear, loglinear, and negative exponential. In terms of face validity (signs and magnitudes of estimated coefficients), goodness of fit (F), and stability over time (similarity of results for main and validation samples), the three distributed lag brand share models below, especially the negative exponential model, were the most successful.

$$B_t = \beta_1 + \beta_2 B_{t-1} + \beta_3 A_t^1 + \beta_4 A_t^2 + \beta_5 A_t^3 + \beta_6 A_t^4 + \beta_7(\bar{p}_t - p_t) + \beta_8(\bar{p}_{t-1} - p_{t-1}) + \beta_9(\bar{d}_t - d_t) + \beta_{10}(\bar{d}_{t-1} - d_t) + \epsilon \quad (1)$$

$$B_t = e^{\beta_1}(B_{t-1})^{\beta_2}(A_t^1)^{\beta_3}(A_t^2)^{\beta_4}(A_t^3)^{\beta_5}(A_t^4)^{\beta_6}$$

$$\cdot (p_t/\bar{p}_t)^{\beta_7}(p_{t-1}/\bar{p}_{t-1})^{\beta_8}(d_t/\bar{d}_t)^{\beta_9}(d_{t-1}/\bar{d}_{t-1})^{\beta_{10}} \cdot e^{\epsilon} \tag{2}$$

$$B_t = 1 - (1 - B_{t-1})^{\beta_2} \exp\,[-\beta_1 + \beta_3 A_t{}^1 + \beta_4 A_t{}^2 + \beta_5 A_t{}^3 + \beta_6 A_t{}^4 + \beta_7(\bar{p}_t - p_t) + \beta_8(\bar{p}_t - p_t) + \beta_9(\bar{d}_t - d_t) + \beta_{10}(\bar{d}_t - d_t) + \epsilon] \tag{3}$$

where:

B_t = share of brand 1 in week t,
$= S_t/T_t$,
S_t = sales of brand 1 in week t,
T_t = sales of coffee in week t,
$A_t{}^m$ = measure of brand 1 advertising outlay in medium m in week t, where:

m	*medium*
1	Network television
2	Spot television
3	Magazines
4	Newspapers

p_t = mean price for purchases of brand 1 not made under deal in week t,
$\bar{p}_t$ = mean price for purchases of coffee in week t (excluding brand 1),
d_t = mean price for purchases of brand 1 made under deal in week t, and
$\bar{d}_t$ = mean price for purchases of coffee made under deal in week t (excluding brand 1).

FINDINGS

The parameters of these 3 models were estimated with the 80-week main sample (weeks 27–106). Ordinary least squares estimates were made for all families taken together and for each of the four clusters defined by each of the three clustering methods (naive, two-variable, and four-variable). For the regressions for a given cluster, the values of the price and deal variables were based on the purchases of only those families in the cluster.

As shown in Table 2, statistically significant F-values occurred more often with the two- and four-variable

Table 3

COEFFICIENT ESTIMATES FOR LINEAR MODEL

Cluster method		*Expected sign*	+	−	+	+	+	+	+	−	+	−	R^2	F
			Cluster											
			b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}		
All families			.09	−.06	−.87	4.38	3.60[b]	.05	.36[a]	.02	−.05	.09	.31[a]	3.39[a]
				(.13)	(.50)	(3.20)	(1.82)	(.05)	(.11)	(.11)	(.07)	(.07)		
Naive	1		.06	−.13	−1.22	8.77	2.56	.11	.15	.07	.09	.02	.17	1.61
				(.13)	(.85)	(5.53)	(2.92)	(.08)	(.10)	(.10)	(.09)	(.10)		
	2		.18	−.01	.56	−4.77	6.09	−.11	.42[b]	−.19	.01	−.02	.07	.55
				(.12)	(2.42)	(12.63)	(7.63)	(.21)	(.22)	(.23)	(.16)	(.14)		
	3		.03	−.01	−.36	1.85	3.02[b]	.02	.07[b]	−.02	−.02	.03	.20	1.87
				(.13)	(.40)	(2.56)	(1.48)	(.04)	(.04)	(.04)	(.05)	(.06)		
	4		.28	−.16	−1.51	5.54	6.62	.04	.80[a]	.43[b]	.01	−.06	.33[a]	3.75[a]
				(.12)	(1.04)	(6.64)	(3.96)	(.10)	(.20)	(.22)	(.13)	(.13)		
Two-variable	1		.05	−.13	−.49	2.83	3.95[b]	.02	.13[b]	−.00	.02	.06	.22[a]	2.21[a]
				(.13)	(.45)	(2.92)	(1.60)	(.04)	(.06)	(.06)	(.06)	(.06)		
	2		.14	−.16	.61	4.76	6.49	−.05	.13	.07	−.30[b]	.21[b]	.14	1.21
				(.12)	(2.34)	(9.33)	(5.73)	(.16)	(.10)	(.10)	(.14)	(.11)		
	3		.08	.05	.84	1.82	2.30	.01	.40	.03	.14[b]	−.06	.51[a]	8.03[a]
				(.12)	(.60)	(3.98)	(2.31)	(.05)	(.06)	(.08)	(.07)	(.07)		
	4		.49	.08	−4.86[b]	5.94	5.29	.15	1.80[a]	.66	.06	−.18	.29[a]	3.11[a]
				(.12)	(2.43)	(14.72)	(8.15)	(.22)	(.58)	(.61)	(.34)	(.35)		
Four-variable	1		.08	−.20	−.28	−.93	3.45	.01	.07[b]	−.05	.10	−.05	.18	1.64
				(.12)	(.53)	(3.30)	(1.91)	(.05)	(.04)	(.04)	(.07)	(.07)		
	2		.05	.11	−.80	2.03	3.16	.07	.18[b]	−.08	−.09	−.05	.26[a]	2.68[a]
				(.12)	(.62)	(3.84)	(2.27)	(.06)	(.07)	(.07)	(.09)	(.08)		
	3		.07	−.07	−.74	6.90	6.20	−.01	.39[a]	.12	−.05	.07	.29[a]	3.08[a]
				(.13)	(1.06)	(6.86)	(3.65)	(.09)	(.12)	(.13)	(.06)	(.06)		
	4		.21	−.06	−.64	7.14	.29	.08	1.20[a]	−.03	.11	−.20[b]	.38[a]	4.66[a]
				(.12)	(1.00)	(6.46)	(3.76)	(.10)	(.22)	(.24)	(.12)	(.11)		

Standard errors in parentheses.
[a] Significant at .01 level.
[b] Significant at .05 level.

defined clusters than with the naively defined clusters. Moreover, for a given clustering method, the loglinear models appeared inferior to the linear and negative exponential models.

Estimates of Marketing Policy Effects

Generally the results for the linear and negative exponential models were quite similar. For that reason and for brevity, the regression coefficients for the linear model alone are described in detail (Table 3). If the selected purchase variables are a useful basis for segmentation, then the coefficient estimates for each cluster should differ. The statistical significance of such differences were examined with an F-test described by Johnston [4, pp. 136–7] and expanded by Beckwith [1]. As shown in Table 4, the set of coefficient estimates did significantly differ across all four clusters, regardless of clustering methods. Moreover, in general the coefficient estimates between clusters of families with similar usage rates but different loyalty rates, and similar loyalty rates but different usage rates also were significantly different.

It was expected that the effects of the price and deal variables would be relatively larger on the heavier usage clusters. The price coefficient was the coefficient estimate most often significant in the regressions, and it was generally larger for the heavier usage clusters. The price and deal coefficients as a group did significantly differ across all clusters for the two- and four-variable clusters; but in the pairwise comparisons, they significantly differed for only one of the light/heavy user pairs (Table 4). Significant differences for the price and deal coefficients among the four-variable clusters were likely not found because clusters 2, 3, and 4 all consisted of heavy users.

On the other hand, for the two- and four-variable clusters, the advertising coefficients were found to significantly differ for the three matches possible between clusters of families having similar usage rates and brand loyalties rated as either low or high (1 vs. 2 and 3 vs. 4). Moreover, the coefficients of the advertising variables as a group did significantly differ across all clusters for all three clustering methods. Individually the advertising coefficients were disappointing—only two were significant and one of these (for the naive clusters) was of the wrong magnitude.

Sales Estimates

For each cluster, each of the 3 regression models was employed to estimate sales in that cluster for each week in the 80-week main sample. These weekly sales estimates were then aggregated across clusters to yield weekly sales estimates for the total sample of families. Finally, these total sales estimates were correlated with the actual sales observed during each week, thereby providing a measure of the overall goodness of fit of each clustering method and model. Benchmarks were obtained by correlating estimated and actual sales from the regressions based on all the families together.

Table 4

F-VALUES, TESTS FOR EQUALITY OF COEFFICIENTS OF LINEAR MODEL

Cluster method	*Clusters compared*	*Coefficients compared*		
		All	*Advertising*	*Price and deal*
Naive	1, 2, 3 & 4	9.45[a]	5.10[a]	1.64
	1 & 2	2.27[a]	1.13	.49
	3 & 4	5.45[a]	2.76[a]	1.14
	1 & 3	2.00[b]	1.10	.11
	2 & 4	.88	.55	.17
Two-variable	1, 2, 3 & 4	12.48[a]	6.85[a]	4.48[a]
	1 & 2	4.45[a]	2.64[a]	.56
	3 & 4	3.96[a]	2.29[a]	1.11
	1 & 3	5.07[a]	2.38[a]	1.23
	2 & 4	6.18[a]	.25	1.83[b]
Four-variable	1, 2, 3 & 4	17.20[a]	6.10[a]	4.35[a]
	1 & 2	.93	.21	.43
	3 & 4	6.94[a]	2.70[a]	1.43
	1 & 3	1.96[b]	.27	1.04
	2 & 4	7.78[a]	2.78[a]	1.53

[a] Significant at .01 level.
[b] Significant at .05 level.

Table 5 shows the linear and negative exponential models to be much superior to the loglinear model with respect to *goodness of fit* (correlation with the main sample data). The similarity of results for the linear and negative exponential models suggests the brand studied is on the relatively straight portion of the negative exponential function—an expected result for a mature brand such as the one studied.

The use of any clustering method and a linear or negative exponential model produced a higher R^2-value than that obtained by estimating a single function for all families. Interestingly, in this application the most complex clustering method—that based on four variables—resulted in the closest fit to actual weekly sales.

A key test of any predictive procedure should be how well it predicts. The predictive abilities of these models were evaluated by employing the regression coefficients based on the main sample to estimate sales for each week of the 24-week validation sample. (Recall these observations had not been previously analyzed.) Sales were estimated for each cluster and aggregated to produce sales estimates for the total sample; then these total sales estimates were correlated with actual weekly sales. The results (Table 5) show the predictive ability of the all-family models to be superior to the models based on the naive and two-variable cluster approaches. However, the predictive performance of the linear and negative exponential models based on the four-variable clusters —the two models that provided the best fit to the main sample observations—resulted in the most accurate

Table 5
R^2-VALUES FOR ESTIMATED S. ACTUAL SALES

Clustering method	*Main sample*			*Validation sample*		
	Linear	*Log-linear*	*Negative-exponential*	*Linear*	*Log-linear*	*Negative-exponential*
All families	.31	.28	.31	.29	.32	.29
Naive	.40	.27	.39	.23	.21	.23
Two-variable	.41	.24	.41	.22	.11	.24
Four-variable	.47	.30	.46	.34	.24	.34

predictions of sales during the 24-week validation interval.

SUMMARY AND CONCLUSIONS

A methodology for employing cluster analysis to segment markets for the purpose of estimating marketing policy effects has been described. In brief, it consists of initially clustering sampling units (buyers, geographical areas, etc.) on the basis of variables a priori expected to be related to their sensitivities to various marketing policies. Next, response functions are estimated for each cluster or segment. Aggregating the estimates from these segment response curves produces the estimate for the entire market. This approach may have utility if nonlinear response functions or dependent variables other than sales are considered, or if exogenous variables that differ by segment are employed.

The approach was illustrated by examining the effects of the marketing policies of a particular brand of regular coffee. Purchase variables were used to identify segments exhibiting significantly different responses to advertising and price. While all the segment-by-segment approaches considered did not result in improved accuracy of prediction, those models providing the best fits to an 80-week period, when extrapolated to 24 later weeks, resulted in more accurate predictions than approaches that ignored segment differences.

REFERENCES

1. Beckwith, Neil. "Test of Equality Between Coefficients in Several Linear Regressions," working paper, Columbia University.
2. Frank, Ronald E. and Paul E. Green. "Numerical Taxonomy in Marketing Analysis; A Review Article," *Journal of Marketing Research*, 5 (February 1968), 83–93.
3. Frank, Ronald E. and William Massy. "Estimating the Effects of Short-Term Promotional Strategy in Selected Market Segments," in Patrick J. Robinson, ed., *Sales Promotion Analysis: Some Applications of Quantitative Techniques*. Boston: Allyn and Bacon, 1967.
4. Johnston, J. *Econometric Methods*. New York: McGraw-Hill, 1963.
5. Sexton, Donald E., Jr. "Estimation of Marketing Policy Effects on Sales," *Journal of Marketing Research*, 7 (August 1970), 338–47.
6. ———. "A Microeconomic Model of the Effects of Advertising," *Journal of Business*, 45 (January 1972), 29–41.
7. ———. "Overspending on Advertising," *Journal of Advertising Research*, 11 (December 1971), 19–25.