SACHIN GUPTA and PRADEEP K. CHINTAGUNTA

The authors propose an extension of the logit-mixture model that defines prior segment membership probabilities as functions of concomitant (demographic) variables. Using this approach it is possible to describe how membership in each of the segments, segments being characterized by a specific profile of brand preferences and marketing variable sensitivities, is related to household demographic characteristics. An empirical application of the methodology is provided using A.C. Nielsen scanner panel data on catsup. The authors provide a comparison with the results obtained using the extant methodology in estimation and validation samples of households.

On Using Demographic Variables to Determine Segment Membership in Logit Mixture Models

The problems of identifying segments of households in a population, determining their associated sensitivities to marketing variables, and investigating the possible bases of segmentation all have occupied the attention of marketing researchers over the last two to three decades (see Wind 1978 for an exhaustive review). Applications of latent class models (Lazarsfeld and Henry 1968), which use aggregate brand switching matrices to uncover the underlying segments in a given market, have been well documented in the marketing literature (Grover and Srinivasan 1987; Jain, Bass, and Chen 1990).

Kamakura and Russell (1989) develop a multinomial logit-mixture model for market segmentation that is based on differences in preferences and price sensitivities across households. This more recent innovation in the latent class paradigm enables us to identify the number of segments and their respective sizes and also determine the sensitivity of households in each segment to marketing variables such as price (see also Dunn, Reader, and Wrigley 1987). The rationale underlying this approach is as follows: There is a fixed, finite number of segments in the market. Households belong to each of these segments with some probability. These probabilities are assumed to be, *a priori*, invariant across households and, therefore, are interpreted as representing segment sizes. Conditional on membership in a particular segment, the probability of a household choosing a brand is given by the logit model. The intrinsic brand preferences (i.e., the intercepts of the logit model) as well as sensitivity to marketing variables are allowed to vary across segments. This enables researchers to characterize segments by their preferred brand and sensitivities to price and promotional variables. A household then is assigned to a segment by updating the household invariant segment membership probabilities by the purchase history of that household (using an empirical Bayes procedure) and picking the segment for which the household has the largest posterior probability of membership.

For a segmentation scheme to be actionable, that is, for it to be used by marketing managers for purposes of targeting and positioning, it is necessary to characterize segments by demographic variables, data for which are readily available. We provide an illustrative example of a situation in which information on demographic characteristics is necessary to implement segment-specific marketing activities.

In a recent discussion with a major East Coast direct marketing company, the following issue was raised: Given a mailing list of households who are potential buyers of the variety of products marketed by the company, how could the company target promotional materials for specific products to only those households that were more likely to purchase those products? The only household-

Sachin Gupta is an Assistant Professor of Marketing, J.L. Kellogg Graduate School of Management, Northwestern University. Pradeep K. Chintagunta is an Assistant Professor of Marketing, S.C. Johnson Graduate School of Management, Cornell University. The authors are grateful to the editor and three anonymous reviewers for their numerous helpful comments and suggestions.

specific information available to the direct marketing company are demographic characteristics such as income, household size, education, ethnic background. Note that the objective in this case is for the company to direct its actions at specific *households*. Demographic information at the level of the actionable unit, that is, the household, is required for the implementation of segment specific marketing strategies.

This example highlights the need to relate explicitly segment membership to observable demographic information and be able to classify households into segments on the basis of their demographic characteristics. In the context of the Kamakura and Russell (1989) multinomial logit-mixture model, this requires being able to explain the probability of a household belonging to a particular segment by that household's demographic characteristics. Dayton and Macready (1988) provide a new type of latent class model in which the probability of latent class membership is functionally related to concomitant variables. These concomitant variables include demographic and other consumer-specific characteristics that assume known, fixed values. The relationship between the concomitant variable latent class models (Dayton and Macready 1988) and the multinomial logit-mixture model (Kamakura and Russell 1989) is the following: If the logit probability of segment membership in the multinomial logit-mixture model is made an explicit function of concomitant (demographic) variables, and if the latent class probabilities and the membership probabilities of the Dayton and Macready (1988) model are logit, then the resulting models are equivalent.

We provide such an extension of the multinomial logit-mixture model using demographic variables to explain segment membership probabilities. Hence, the proposed approach can be viewed as a logical combination of the Kamakura and Russell (1989) multinomial logit-mixture model and the Dayton and Macready (1988) concomitant variable latent class model.

The proposed approach has both methodological and substantive merits. From a methodological perspective, (1) if demographic variables do influence segment membership, the proposed approach incorporates this information in the estimation of the parameters of the multinomial logit-mixture model and (2) it enables us to test which of the demographic variables are useful predictors of segment membership.

In a substantive sense, the approach enables implementation of an actionable segmentation scheme such as that in the illustration discussed previously. Using the proposed approach, it is possible to describe how membership in each segment, with the segments characterized by a specific profile of brand preferences and marketing variable sensitivities, is related to observable household characteristics. Furthermore, establishing such a relationship would enable us to classify households into segments on the basis of demographic information alone, which is the objective of most segmentation schemes (Wind 1978). Classification of households on the basis of posterior segment membership probabilities obtained from the conventional multinomial logit-mixture model (as in Kamakura and Russell 1989) requires information on the purchase histories of households. Such information may not be available for households outside the estimation sample, though information regarding individual shoppers' purchase histories is becoming increasingly available through supermarket purchase information systems. If such information is available, an updating procedure such as that proposed by Kamakura and Russell (1989) can be applied to the household-specific (prior) probabilities of segment membership obtained using the methodology proposed here.

Bucklin and Gupta (1992) estimate a nested multinomial logit-mixture model to identify response segments in the liquid laundry detergent market. To relate segment membership to observable demographic characteristics, they regress the posterior segment membership probabilities on those demographic variables. Their "exploratory analysis" (Bucklin and Gupta 1992, p. 212) appears to be intended primarily to characterize the segments on the basis of demographics rather than for purposes of classifying households into response segments. In contrast, the methodology proposed here in which *a priori* segment membership probabilities are made functions of demographic variables, can be used for both characterizing segments and assigning households to preference and response segments on the basis of their demographic characteristics. However, it must be noted that in certain applications the approach suggested by Bucklin and Gupta (1992) could offer certain advantages as compared with the approach proposed here. Specifically, when there are a large number of demographic variables, testing the effects of multiple combinations of these variables on segment membership probabilities could be computationally less cumbersome using the regression approach of Bucklin and Gupta (1992).

We provide an empirical application of the proposed methodology using the A.C. Nielsen scanner panel data on the purchases of catsup. Our results indicate that demographic variables such as income and household size significantly affect the segment membership probabilities, and hence this information should not be ignored when estimating the multinomial logit-mixture model. We find that low-income households tend to be price- and promotion-sensitive, whereas larger households prefer the larger brand sizes. The proposed methodology enables us to *characterize* segments (both qualitatively and quantitatively) and also *classify* households to segments on the basis of demographic variables. The validity of the methodology is then assessed in a hold-out sample of households.

In the next section, we describe the model formulation, followed by a section containing the empirical application. We present conclusions and directions for further research in the final section.

MODEL FORMULATION

We begin by assuming that there exist S segments in the market under investigation ($s = 1, 2, \ldots, S$). Each segment consists of a (to be estimated) number of households that are assumed to be homogenous with respect to their preferences for brands as well as their sensitivity to marketing variables. Segments, however, differ in both preferences and responsiveness to marketing effort. We now characterize the nature of brand choice behavior in each of the S segments.

Let us define $P_h^t(k|s)$ as the probability that household h chooses brand k on purchase occasion t conditional on the household belonging to segment s. Given the classic random utility framework and the assumption of iid extreme value distributions for the stochastic component of a household's utility for a brand, this probability can be written as

$$(1) \qquad P_h^t(k|s) = \frac{\exp(\tilde{\theta}_{ks} + \beta_s \bar{X}_{hk}^t)}{\sum_{\ell=1}^{K} \exp(\tilde{\theta}_{\ell s} + \beta_s \bar{X}_{h\ell}^t)},$$

where $\tilde{\theta}_{ks}$, $k = 1, \ldots, K$ are the intrinsic preferences associated with the K brands by households in segment s and β_s is the parameter vector associated with the vector of marketing (explanatory) variables $\bar{X}_{h\ell}^t$. Normalizing equation 1 with respect to brand K, we obtain

$$(2) \qquad P_h^t(k|s) = \frac{\exp(\theta_{ks} + \beta_s X_{hk}^t)}{1 + \sum_{\ell=1}^{K-1} \exp(\theta\ell_s + \beta_s X_{h\ell}^t)},$$

where $\theta_{ks} = \tilde{\theta}_{ks} - \tilde{\theta}_{Ks}$ and $X_{hk}^t = \bar{X}_{hk}^t - \bar{X}_{hK}^t$, $k = 1, 2, \ldots, K - 1$.

The probability that a household h ($h = 1, 2, \ldots, H$) belongs to segment s, P_{hs} depends on the vector of demographic variables D_h specific to that household. This probability can be represented as

$$(3) \qquad P_{hs} = \frac{\exp(\bar{\alpha}_s + \bar{\gamma}_s D_h)}{\sum_{r=1}^{S} \exp(\bar{\alpha}_r + \bar{\gamma}_r D_h)},$$

where $\bar{\alpha}_s$ is the intercept and $\bar{\gamma}_s (s = 1, 2, \ldots, S)$ are unknown segment-specific parameters to be estimated and denote the contribution of the various demographic variables to the probability of segment membership. Equation 3 is the logit probability of household h belonging to segment s. Clearly, one could choose functional forms other than the logit to represent the probability of segment membership, the only constraints being that $\sum_{s=1}^{S} P_{hs} = 1$ and $0 \leq P_{hs} \leq 1$. However, given the ease in interpreting the parameters $\bar{\alpha}_s$ and $\bar{\gamma}_s$ from a logit specification as well as in model estimation, we resort to the formulation in equation 3. Normalizing the right-hand side of this expression with respect to the parameters of segment S gives us

$$(4) \qquad P_{hs} = \frac{\exp(\alpha_s + \gamma_s D_h)}{1 + \sum_{r=1}^{S-1} \exp(\alpha_r + \gamma_r D_h)}$$

where $\alpha_s = \bar{\alpha}_s - \bar{\alpha}_S$; $\gamma_s = \bar{\gamma}_s - \bar{\gamma}_S$.[1] The parameter γ_s is the difference in the effect of demographic variable D_h on the probability of membership in segment s from the effect of that variable on the probability of belonging to segment S.

From equation 4, we note that given estimates for α_s and γ_s, $s = 1, 2, \ldots, S - 1$, it is possible to compute a probability of membership for household h in each of the S segments in the market if the household's demographic characteristics are known. Hence, on the basis of an assignment rule such as "membership in the segment with highest probability," it is possible to uniquely assign households to segments with differential sensitivity to marketing variables. In any given product market the information typically available to the manager is the demographic variables characterizing the population of interest. The methodology suggested here, therefore, enables the manager to exploit this information to classify households into groups on the basis of their responsiveness to their own and competing brands' marketing programs. Such an "endogenous" classification has not been possible with the models in the extant literature on logit mixture models. The question that arises immediately is, How do we obtain estimates for α_s and γ_s from available household scanner panel data? We now address this issue.

Given equations 2 and 4, the probability that a randomly selected household h purchases brand k on purchase occasion t, $P_h^t(k)$ is given by

$$(5) \qquad P_h^t(k) = \sum_{s=1}^{S} P_{hs} \cdot P_h^t(k|s),$$

which is the expression for the unconditional probability in a latent class model as shown by Dayton and Macready (1988).

Estimation

In any given panel data application, each household h makes T_h purchases in the category of interest. The probability of occurrence of the sequence of T_h purchases for

[1] Note that alternatively, we could have normalized with respect to the expected values of $\bar{\alpha}_s$ and $\bar{\gamma}_s$ across segments. The parameter estimates and their standard errors for this normalization can be recovered from the estimation of the parameters of equation 4 by using the estimates for α_s and γ_s, $s = 1, 2, \ldots, S - 1$, and the Fisher information matrix corresponding to these estimates. We are grateful to an anonymous reviewer for highlighting this point.

the household conditional on its membership in segment s is given by

$$(6) \qquad P_{h|s} = \prod_{t=1}^{T_h}\prod_{k=1}^{K}[P_h^t(k|s)]^{\delta_{kh}^t},$$

where

$$\delta_{kh}^t = \begin{cases} 1 & \text{if household } h \text{ buys brand } k \text{ on occasion } t \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function for household h, L_h, is given by

$$(7) \qquad L_h = \sum_{s=1}^{S} P_{hs} \cdot P_{h|s}$$

and the sample likelihood function is

$$(8) \qquad L = \prod_{h=1}^{H} L_h.$$

The unknown parameters of the problem, that is, α_s; γ_s, $s = 1.\ 2, \ldots, S - 1$; θ_{ks}, $k = 1, 2, \ldots, K - 1$; β_s, $s = 1, 2, \ldots, S$ then are estimated using the method of maximum likelihood. The number of segments S is determined by carrying out the estimation for 2, 3, 4, 5, ... segments until there is no significant improvement in model fit between S and the $S + 1$ segment solutions. The difference in model fit is evaluated using the Bayesian Information Criterion (BIC; see Allenby 1990).

EMPIRICAL APPLICATION

Data

The A.C. Nielsen scanner panel data on the purchases of catsup were used for the empirical analysis. Of a total of 2500 households in the panel, 945 households, accounting for 7575 purchases of the six largest brand-sizes over the two-year period of the data, were singled out for the analysis.[2] These six brand-sizes were Heinz (28 oz., 32 oz., 40 oz., and 64 oz.), Hunt's (32 oz.), and Del Monte (32 oz.) and accounted for 85% of the total market. Of the 945 households, 709, accounting for 5611 purchases, were used for purposes of model estimation, and the remaining 236 households constituted the validation sample. In Table 1, we provide the descriptive statistics for the data.

From Table 1 we note the Heinz 28 and 32 oz. sizes have the largest share of purchases and promotions. The primary difference between these two brand-sizes is that the 32 oz. pack comes in a glass bottle whereas the 28 oz. is a squeezable plastic bottle. Table 1 also provides the average values of the demographic variables, income, household size, and average age of heads of household for the estimation and validation samples. Besides these, two other variables—education and mean expenditure levels—were tried but dropped from the final model because they did not result in a significant improvement in model fit.

Results

Table 2a provides the parameter estimates and the t-ratios for the two model formulations of interest. The

[2] These households were selected using the criterion that at least 75% of their purchases during the two-year period were from among the six brand-sizes.

Table 1
DESCRIPTIVE STATISTICS FOR CATSUP DATA

1a Brand Data

Brand	*Share (%)*	*Average Price (¢/oz.)*	*Display*[a]	*Feature*[b]
Heinz 64 oz.	3.9	4.59	.7	2.1
Heinz 40 oz.	7.6	4.80	2.5	2.2
Heinz 32 oz.	43.6	3.33	5.9	10.7
Heinz 28 oz.	25.6	4.50	6.1	7.2
Hunts 32 oz.	13.8	3.47	5.6	5.6
Del Monte 32 oz.	5.5	3.49	4.3	5.6

1b Demographic Data[c]

Variable	*Estimation Sample Mean*	*Validation Sample Mean*
Income	6.59	6.71
Household size	3.16	3.23
Average age of heads of household	48.06	48.68
Number of households	709	236

[a,b] Percentage of purchases made on Display and Feature respectively.
[c] Income is coded as 1 = less than $5,000; 2 = $5,000–$10,000; 3 = $10,000–$15,000, and so on.

Table 2a
PARAMETER ESTIMATES AND (t-RATIOS)
THREE-SEGMENT SOLUTION[a]

	Without Demographics			*With Demographics*		
Variable	*Segment 1*	*Segment 2*	*Segment 3*	*Segment 1*	*Segment 2*	*Segment 3*
Heinz 64	−.403	6.570	.797	−.394	6.522	.759
	(−1.345)	(6.414)	(1.468)	(−1.320)	(6.398)	(1.344)
Heinz 40	1.877	5.877	4.395	1.861	5.830	4.386
	(12.406)	(5.681)	(13.427)	(12.183)	(5.663)	(13.481)
Heinz 32	2.380	4.567	2.654	2.377	4.558	2.647
	(30.164)	(4.541)	(8.711)	(30.079)	(4.536)	(8.705)
Heinz 28	2.709	4.948	5.377	2.714	4.924	5.363
	(22.639)	(4.825)	(16.559)	(22.547)	(4.821)	(16.680)
Hunts 32	.964	2.705	2.231	.957	2.659	2.256
	(10.593)	(2.622)	(7.084)	(10.254)	(2.571)	(7.147)
Price	−1.898	−1.057	−1.242	−1.899	−1.043	−1.231
	(−31.465)	(−6.461)	(−11.300)	(−31.856)	(−6.453)	(−11.375)
Display	1.062	.831	1.027	1.068	.843	1.043
	(9.489)	(2.900)	(5.792)	(9.564)	(2.929)	(5.906)
Feature	1.171	.588	.425	1.165	.607	.434
	(13.494)	(2.469)	(2.475)	(13.353)	(2.559)	(2.558)
Segment size	.653	.080	.267	.621	.135	.244
−LL		5649			5630	
ρ^2		.438			.440	
$\bar{\rho}^2$		.440			.443	
Number of parameters		26			32	
Likelihood ratio statistic					38[b]	

[a] Del Monte 32 was treated as the base brand.
[b] χ^2-value significant at the 5% level of significance.

first model is the one used by Kamakura and Russell (1989) and Bucklin and Gupta (1992) in which *a priori* segment membership probabilities are assumed invariant across households (the "without-demographics model"). The alternative formulation, that is, our proposal, uses information on demographic variables directly in the estimation (the "with-demographics model").

For both the model formulations, we tried 2-, 3- and 4-segment solutions. Because a K-segment model is not strictly nested within the $K + 1$ segment model, we used BIC to determine the appropriate number of segments. We found that a 3-segment solution was the preferred specification for the with- and the without-demographics models. Given a 3-segment solution, we note that the without-demographics model is nested within the with-demographics model. Using a likelihood-ratio test of model log-likelihoods, we find that the with-demographics model fits the data significantly better than the without-demographics model at the 5% level of significance, but the improvement in fit ($\bar{\rho}^2$) is very slight (an increase of .003 in the value of $\bar{\rho}^2$).

From Table 2a we also note that the parameter estimates from the two specifications are comparable in magnitude. Specifically, segments 1 and 3 have the highest preference for Heinz 28 oz., whereas segment 2's preference is for Heinz 64 oz. Further, segment 1 appears to be the most price- as well as promotion-sensitive. The interesting point of difference between the two formulations, however, is the disparity in segment sizes. Whereas segment 1 is of about equal magnitude, the model

Table 2b
PARAMETER ESTIMATES AND (t-RATIOS)
EFECTS OF DEMOGRAPHIC VARIABLES ON SEGMENT MEMBERSHIP PROBABILITIES[a]

	Without Demographics		*With Demographics*	
Variable	*Segment 1*	*Segment 2*	*Segment 1*	*Segment 2*
Intercept	.893[b]	−1.214[b]	.886	−1.070
	(6.514)	(−6.525)	(1.400)	(−.984)
Income			−.168[b]	−.082
			(−4.535)	(−1.328)
Household size			.167[c]	.356[b]
			(1.810)	(2.530)
Age			.014	−.016
			(1.482)	(−.933)

[a] Segment 3 is treated as the "base" segment.
[b] Significant at the 5% level of significance.
[c] Significant at the 10% level of significance.

with demographic variables predicts a larger segment 2 than does the restricted model.

Table 2b provides the parameter estimates for the effects (as defined in the previous section) of demographic variables on segment membership probabilities using the proposed approach. Because segment 3 is the base segment, all its parameter estimates are normalized to zero. We note that only the effects of income and household size are significant based on this normalization. The negative and significant effect of income on segment 1 membership relative to membership in segment 3 indicates that a low income implies a higher probability of belonging to segment 1. This makes intuitive sense, because from Table 2a we see that households in segment 1 have the highest sensitivity to price and promotional variables.

Turning to segment 2, we note that the only significant variable determining membership in this segment, relative to segment 3, is household size. Furthermore, the magnitude of the corresponding parameter is larger than that for segments 1 and 3. This implies that a larger household has a greater probability of belonging to segment 2 than to any other segment. Though segment 2 is characterized by the lowest levels of price and display sensitivity, it is interesting to note that households in this segment have the largest values of intrinsic preference for Heinz 64 oz. and 40 oz. brand-sizes. This preference of large households for the large-sized packs seems to be reasonable intuitively. Therefore, differences in segment level preferences as well as responsiveness to marketing activities can be directly linked to demographic variables using the methodology proposed here.

Characterization of segments and Comparison with Alternative Approaches

To characterize the nature of households belonging to the different segments, it is first necessary to assign households to these segments. To implement a segmentation scheme in practice, a manager may have information about only the demographic characteristics of households or may have information on purchase histories of these households as well. Depending on the nature of available information, one of two possible assignment schemes can be adopted:

1. When only demographic information is available, the parameter estimates in Table 2b and the expression in equation 4 can be used to compute the segment membership probabilities for each household. Households then can be assigned to segments on the basis of the "largest probability" rule.[3]
2. When the additional information on purchase histories is available, the first step is to compute the probabilities as in (1). These household-specific probabilities are then treated as priors that are updated using the purchase history information by means of an empirical Bayes procedure. Households are then assigned to segments on the basis of the posterior probabilities using the "largest probability" rule.

We present results only for method 2 here. We compare the results obtained with those from an extant alternative approach. In this latter approach, the segment sizes presented in Table 2a from the without-demographics model are treated as household invariant prior probabilities. These probabilities then are updated using the purchase histories of households as described in method 2. Such an approach has been proposed by Kamakura and Russell (1989).[4]

We report in Table 3 the means and standard deviations of the income and household size of households assigned to each segment along with the segment sizes. Note that both models delineate the three segments consistent with our previous discussion on the effects of demographic variables on segment membership. Specifically, segment 1 is characterized by low income and medium household size; segment 2 by medium/high income and large household size; and segment 3 by high income and small household size. Though the differences in mean demographic profiles are small, these qualitative descriptions of the demographic profiles of the segments can be used for the implementation of segmentation schemes by managers.

Table 3 also highlights the key differences between

Table 3
SEGMENT CHARACTERIZATION AND COMPARISON FOR ESTIMATION SAMPLE:
MEAN (STANDARD DEVIATION) OF DEMOGRAPHIC VARIABLES

	Segment 1	*Segment 2*	*Segment 3*
Income			
With[a]	6.15	7.26	7.58
	(2.74)	(2.83)	(3.07)
Without[b]	6.34	7.00	7.38
	(2.85)	(2.80)	(2.99)
Household size			
With	3.15	3.63	3.04
	(1.29)	(1.10)	(1.27)
Without	3.13	3.59	3.12
	(1.30)	(1.11)	(1.27)
Proportion of households assigned			
With	63	9	28
Without	67	8	25

[a] Refers to "with demographics" model.
[b] Refers to "without demographics" model.

[3] Note that such an approach could capitalize on chance and therefore could show a difference in segment-level demographics even if there were no true differences.

[4] Results obtained using method (1) are available from the authors on request.

the with- and without-demographics models. First, note that the two models assign different numbers of households to the three segments.[5] Furthermore, the with-demographics model leads to a larger difference in the mean income levels of segments 1 and 3. Hence, these two segments are more clearly delineated on the income variable. This finding is consistent with the results in Table 2b, in which we find that the income variable has a negative and significant effect on membership in segment 1 relative to segment 3. A similar result is obtained for the household size variable. Specifically, we find that households in segments 1 and 2 are delineated better from households in segment 3 using the with-demographics specification. This is once again consistent with the statistically significant parameter estimates for the household size variable in Table 2b. We reiterate that though differences exist between the two model specifications in the characterization of segments, these differences are small because of the marginal improvement in fit obtained by the inclusion of demographic variables.

Yet another approach to using demographic information to characterize segments is by interpreting the parameter estimates obtained by regressing the posterior segment membership probabilities ($\tilde{P}_{hs}$, $s = 1, 2, \ldots, S$) from the without-demographics model on household demographic variables. To ensure comparability of the parameters with those in equation 4, the following regression equations are used for estimation purposes[6]:

$$\ell n\left(\frac{\tilde{P}_{hs}}{\tilde{P}_{hS}}\right) = \alpha_s + \gamma_s D_h + \epsilon_{hs}; \quad s = 1, 2, \ldots, S - 1.$$

In our empirical application, $S = 3$ and hence, we have two regression equations.[7] Because the error terms in these regressions are correlated, using seemingly unrelated regression (SUR) methods should improve efficiency of the estimates (Zellner 1962). However, because the predictor matrices (i.e., demographic variables D_h) ar identical for the two equations, SUR estimates are identical to those obtained from ordinary least squares (OLS estimation. In Table 4 we present the parameter estimates and their t-ratios for these regressions.[8]

Table 4
REGRESSION OF POSTERIOR PROBABILITIES ON DEMOGRAPHIC VARIABLES: PARAMETER ESTIMATES AND (*t*-RATIOS)

Dependent Variable $= \ell n\left(\frac{\tilde{P}_{hs}}{\tilde{P}_{hS}}\right)$, $s = 1, 2$

Variable	*Segment 1*	*Segment 2*
Intercept	.19	−4.03
	(.10)	(−2.13)
Income	−.30	−.15
	(−2.64)	(−1.32)
Household size	−.51	−.51
	(−1.73)	(−1.77)
Age	−.02	−.03
	(−.62)	(−1.03)
R^2	.047	.024
$\bar{R}^2$	.035	.012
Number of households	709	709

Note from Table 4 that the highest value of $\bar{R}^2$ is .03 (segment 1). This indicates that the demographic variables do not seem to explain much of the variability i the dependent variable. From this regression analysis segments 1, 2, and 3 can be characterized as being lov income − small household size, medium income − smal household size, and high income − large household size respectively. Based on our previous discussion of the pric sensitivities and size preferences of the three segments it is clear that the characterization on the basis of household size obtained here is not intuitively appealing. Furthermore, the estimates in Table 4 also can be used t classify households to the three segments (on the basi of the "largest probability" rule). Such an exercise resulted in all 709 households being assigned to segmer 3. Hence, it is evident that this approach is unsuitabl for purposes of both characterizing segments as well a assigning households to these segments.

Results for Validation Sample

We now examine the performance of the with- an without-demographics models in a validation sampl consisting of 236 households accounting for 1964 pu

[5] Note, however, that the sizes of segments 2 and 3 obtained in Table 3 appear to be closer to those reported for the without-demographics model in Table 2a.

[6] From Equation 4,

$$P_{hs} = \frac{\exp(\alpha_s + \gamma_s D_h)}{1 + \sum_{r=1}^{S-1} \exp(\alpha_r + \gamma_r D_h)}$$

$$P_{hS} = \frac{1}{1 + \sum_{r=1}^{S-1} \exp(\alpha_r + \gamma_r D_h)}.$$

Hence,

$$\frac{P_{hs}}{P_{hS}} = \exp(\alpha_s + \gamma_s D_h), \quad \text{and}$$

$$\ell n\left(\frac{P_{hs}}{P_{hS}}\right) = \alpha_s + \gamma_s D_h, \quad s = 1, 2, \ldots, S - 1.$$

[7] Bucklin and Gupta (1992) also attempt to characterize segments by regressing posterior segment membership probabilities on demographic variables. In their case, however, the dependent variables are $\ell n(\tilde{P}_{hs}/1 - \tilde{P}_{hs})$, $s = 1, 2, \ldots, S$. Hence they have S regressions instead of $S - 1$.

[8] Because the posterior membership probabilities are skewed towa one or zero, the error in these regressions is likely to be heteroskdastic. We recomputed the standard errors using the heteroskedastici consistent covariance matrix of parameter estimates (White 1980 However, conclusions about significance of parameter estimates remained unaffected.

Table 5
SEGMENT CHARACTERIZATION AND COMPARISON FOR VALIDATION SAMPLE
MEAN (STANDARD DEVIATION) OF DEMOGRAPHIC VARIABLES

		Segment 1	*Segment 2*	*Segment 3*
Income	With	6.64	6.70	6.90
		(2.76)	(3.16)	(2.73)
	Without	6.68	6.88	6.75
		(2.74)	(3.37)	(2.76)
Household size	With	3.31	3.50	2.91
		(1.33)	(1.43)	(1.12)
	Without	3.27	3.47	3.05
		(1.33)	(1.28)	(1.20)
Proportion of households assigned	With	66%	9%	25%
	Without	69%	7%	24%

chases of catsup. The validation exercise used parameter estimates reported in Table 2, purchase histories of the validation sample households (for the without-demographics model), and demographic information for these households (for the with-demographics model).

A commonly used measure of predictive performance in validation samples is the predictive log-likelihood. The values obtained for this measure are -1950.26 for the with-demographics model ($\rho^2 = .446$) and -1953.55 for the without-demographics model ($\rho^2 = .445$). Another measure of the predictive ability of a model is the hit rate, that is, the proportion of choices correctly predicted in a hold-out sample. For this exercise, we set aside the last third of purchase occasions of the 236 households in the validation sample as the hold-out period. These households were assigned to the three segments on the basis of the posterior probabilities of segment membership computed using the first two-thirds of the purchase occasions. Posterior probabilities were computed using the priors from both the with- and the without-demographics models. Brand choices of each household for the hold-out period were predicted using both model specifications. The predictions from each model specification were compared with the actual choices. This exercise revealed a hit rate of 62% for both the with- and without-demographics models. The marginal improvement in predictive ability obtained using the measure of predictive log-likelihood and the lack of improvement in the hit rate for the model that includes demographic variables is consistent with our finding in the estimation sample.

The demographic profiles—mean and (standard deviation)—obtained for the three segments are in Table 5. Households are assigned as described for the estimation sample. The results in Table 5 are largely consistent with those reported in Table 3 and reinforce our finding that the profiles obtained from the with-demographics model provide a clearer delineation of the three segments. Using the parameter estimates in Table 4 to classify households into segments on the basis of demographic variables resulted in all 236 households being assigned to segment 3. In contrast, the with-demographics model classifies 66% of households in segment 1, 9% in segment 2, and 25% in segment 3. Hence, the results obtained from the validation sample corroborate our findings from the estimation sample.

MANAGERIAL IMPLICATIONS AND CONCLUSIONS

We have proposed an extension to extant methodologies by directly incorporating the effects of demographic characteristics on segment membership probabilities in logit mixture models for market segmentation. The procedure is easy to implement and provides a direct link between segment-level sensitivity of households to marketing variables and household demographic characteristics.

Though we find that there is a statistically significant improvement in model fit by including demographic characteristics in the estimation, the only variables that appear to be determinants of segment membership are income and household size. Furthermore, there is little or no improvement in predictive ability when demographic variables are included in the analysis. Taken together, these findings seem to imply that demographic variables currently reported in scanner panel data could contain limited information for purposes of investigating brand choice behavior of households. It would be of interest to repeat the exercise in other contexts to ascertain the value of demographic information. This issue we leave for further research.

The methodology proposed here can be applied to other household purchase decisions such as purchase quantity. It also can be applied to model formulations other than the logit, in the context of latent class analysis. Readers interested in results from application of the proposed methodology to the Poisson regression and the multinomial probit may contact the authors.

In implementing a segmentation scheme, a manager may need to characterize segments by their demographic profiles and/or classify households into segments using demographic variables. To accomplish this, the manager may or may not have access to the information on purchase histories of households. Our proposed methodology is applicable in either situation. Here, we have empirically demonstrated that using information on demographic variables along with information on purchase histories leads to a better delineation of segments than the extant methodology that uses information only on the purchase histories of households. We also have demonstrated the superiority of our approach over the methodology that uses the parameter estimates from the regression of posterior probabilities obtained from the without-demographics model on demographic variables.

REFERENCES

Allenby, Greg (1990), "Hypothesis Testing with Scanner Data: The Advantage of Bayesian Methods," *Journal of Marketing Research*, 27 (November), 379–89.

Bucklin, Randolph E. and Sunil Gupta (1992), "Brand Choice, Purchase Incidence and Segmentation: An Integrated Modeling Approach," *Journal of Marketing Research*, 29 (2), 201–15.

Dayton, Mitchell C. and George B. Macready (1988), "Concomitant-Variable Latent-Class Models," *Journal of the American Statistical Society*, 83 (401), 173–9.

Dunn, Richard, Steven Reader, and Neil Wrigley (1987), "A Nonparametric Approach to the Incorporation of Heterogeneity into Repeated Polytomous Choice Models of Urban Shopping Behavior," *Transportation Research* A, 21A (4/5), 327–43.

Grover, Rajiv and V. Srinivasan (1987), "A Simultaneous Approach to Market Segmentation and Market Structuring," *Journal of Marketing Research*, 24 (May), 139–53.

Jain, Dipak C., Frank M. Bass and Yu-Min Chen (1990), "Estimation of Latent Class Models with Heterogeneous Choice Probabilities: An Application to Market Structuring," *Journal of Marketing Research*, 27 (February), 94–101.

Kamakura, Wagner A. and Gary J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, 26 (November), 379–90.

Lazarsfeld, Paul F. and Neil W. Henry (1968), *Latent Structure Analysis*. Boston: Houghton-Mifflin Company.

White, Halbert (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–38.

Wind, Yoram (1978), "Issues and Advances in Segmentation Research," *Journal of Marketing Research*, 15 (August), 317–37.

Zellner, Arnold, (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57, 348–68.