Sawtooth Software

RESEARCH PAPER SERIES

Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results

Bryan K. Orme,
Sawtooth Software, Inc.
and
Michael A. Heft,
Daymon Associated
1999

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results

Bryan K. Orme, Sawtooth Software, Inc.
Michael A. Heft, Daymon Associates

Introduction

Conjoint analysis has been used extensively over the last three decades in marketing, but few published studies have demonstrated that it can predict actual sales. We believe the sparse evidence is not because conjoint cannot predict real world behavior (if that were the case, why would conjoint analysis continue to be so popular?), but that the data are guarded by organizations with no real incentive to publish the results. We report on a study wherein shoppers at grocery stores were given a CBC interview, and the results used to predict actual sales within the same stores.

The second focus of this paper is to demonstrate that capturing heterogeneity (differences in preference between groups or individuals) can improve predictive validity. Traditionally, CBC has been analyzed in the aggregate, by pooling all respondents and developing a summary set of effects (utilities) to reflect the market. Lately, methods that capture heterogeneity by modeling utilities at the group-level (Latent Class) and even at the individual level (Hierarchical Bayes and ICE—Individual Choice Estimation) have become available and been heralded as better models. After showing that Lclass and ICE do a better job at predicting actual sales for our study than aggregate level logit, we'll spend the remainder of this paper investigating why.

The Grocery Store Study Design

Six-hundred respondents were intercepted within five grocery stores and completed a computerized survey programmed using Ci3. The interview facilitators approached every *n*th customer and assisted with running the survey. A randomized choice experiment that included scanned images of the products was programmed into the questionnaire. Three different choice designs, each covering a different product category, were included. Respondents were asked to indicate which product categories they often purchased, and were randomly selected to complete a CBC interview for a product category for which they qualified. For proprietary reasons, we cannot reveal the grocery store chain nor the categories and brands that were studied. We can say that the purpose of the conjoint research was to determine pricing strategy.

The completed interviews per category were as follows:

Category I	246
Category II	205
Category III	149

Three attributes were included in each design:

1 Brand (picture of the package)
2 An unimportant “decoy” attribute
3 Price (a conditional, customized range for each brand)

Each respondent completed 15 choice tasks, and “None” choices were permitted. The “decoy” attribute was not used in modeling, but helped disguise the purpose of our research (pricing) for the respondents.

Four price points were chosen for each brand, such as:

1 25% lower than average price
2 12.5% lower than average price
3 12.5% higher than average price
4 25% higher than average price

Prices were rounded to the nearest 9-cent increment to better reflect the way these products are actually priced on the shelf.

The Value of Good Pricing Information

Pricing is not only a highly sensitive topic in marketing, but also one that has a major profitability impact. In spite of its importance, pricing is a difficult topic for most managers. A recent McKinsey & Company survey asked managers from over 300 North American companies whether they had done any research to measure or predict price elasticity in the previous year (Clancy and Shulman, 1994). Only 15 percent reported doing any kind of primary research. Consistent with those findings, a survey by marketing consultants Clancy and Shulman, found that only about 12 percent of all American companies do any serious pricing research, and one-third of those have no strategy with which to use the research (Clancy and Shulman, 1994).

Pricing in the supermarket environment is even more difficult since the average supermarket is dealing with over 30,000 stock keeping units (SKUs). Pricing decisions must consider cost, marketing strategy and profitability. In the highly competitive consumer package goods arena, supermarkets must also consider the explicit signal that individual "marker items" convey to consumers regarding the general price levels of that store chain. Thus, the price of a very few frequently purchased items can create a consumer price image that will influence future choice of that chain as a regular shopping declaration.

A specific price issue, that of the optimum pricing relationship between different brands, is one of the most frequent questions that consultants to retailers are asked to answer. Our client commissioned this project because of that need. As a bonus, the data were very useful for validating different methods of analysis.

The Validation Sales Data

Actual units sold as recorded by checkout scanners for the last 52 weeks (reported by week) were provided by the client. Here are some details on each category:

Category I:

A food product ranging in price from $1.29 to $2.49. This product was the fastest moving (in terms of unit sales) with about 80 units sold per week in each store. Prices changed throughout the 52-week period, with all brands holding constant for about 8 weeks at a time. When brands went on sale, there was only modest evidence of stocking-up behavior. Three brands were studied in total.

Category II:

A non-food product ranging in price from $4.19 to $9.19. This product was the slowest moving, with about 5 units sold per week in each store. Prices changed throughout the period, with prices holding constant for between 8 to 24 weeks at a time. Only modest evidence of stocking-up behavior was detected. Five brands were studied in total.

Category III:

A food product ranging in price from $1.69 to $3.89. This product sold about 60 units per week in each store. Prices remained constant for most of the weeks, with certain brands going on sale often throughout the year for only a week or two at a time. When brands went on sale, sales volume temporarily and dramatically increased (by a factor of as much as 6x), reflecting stocking-up behavior. Six brands were studied.

Market simulators based on conjoint analysis reflect steady-state, long-range demand. We should not expect conjoint to accurately predict transitional pricing periods for our grocery store categories. Significant stocking-up behavior can occur, and the degree of that behavior depends on factors such as the shelf life, physical size, households' purchase frequency of the product, and how often the product goes on sale.

For each category, different validation scenarios were chosen, with preference given to periods in which prices were held constant for many consecutive weeks. For category III, only one such stable period could be accumulated. Seven other validation scenarios were constructed from the other two categories. Of the 156 total weeks of sales data available to us (52 weeks for three categories), we discarded 45 transitional weeks wherein prices had recently changed. Viewed from the other perspective, we retained 111/156 = 71% of the observations.

Different Methods for Analyzing Choice Data

We developed market simulators using three different methods for modeling CBC data:

1. Aggregate Logit: A single set of utilities summarizes the preferences of the entire sample. Both main-effects and interaction terms can be modeled.

2. Latent Class (LC): The method simultaneously divides the sample into market segments that differ in preferences, and estimates utilities summarizing the preferences for each group. Both main-effects and interaction terms can be modeled.

3. Individual Choice Estimation (ICE): A set of utilities is estimated to reflect the preferences for each individual. The software currently only provides for main-effects.

Another method besides ICE has proven to be useful for estimating individual-level utilities from choice data: Hierarchical Bayes (HB). Though we did not try using HB for our data set, we expect it would have also worked very well given that ICE and HB have been shown to provide very similar results (Huber 1998).

Validation Performance

Five CBC simulators were developed for each product category:

1. Aggregate logit, main-effects only (Logit ME)
2. Aggregate logit, main-effects plus Brand x Price interaction (Logit BxP)
3. Latent Class, main-effects only (LC ME)
4. Latent Class, main-effects plus Brand x Price interaction (LC BxP)
5. ICE, main-effects only (ICE)

The LC simulators used six segments each. ICE utilities were computed using LC vectors as starting points with no additional iterations. The ICE utilities were not calibrated, but used in their raw form. All models employed logit simulations (Model 2), with no external effects and the scale parameter (exponent) set to unity.

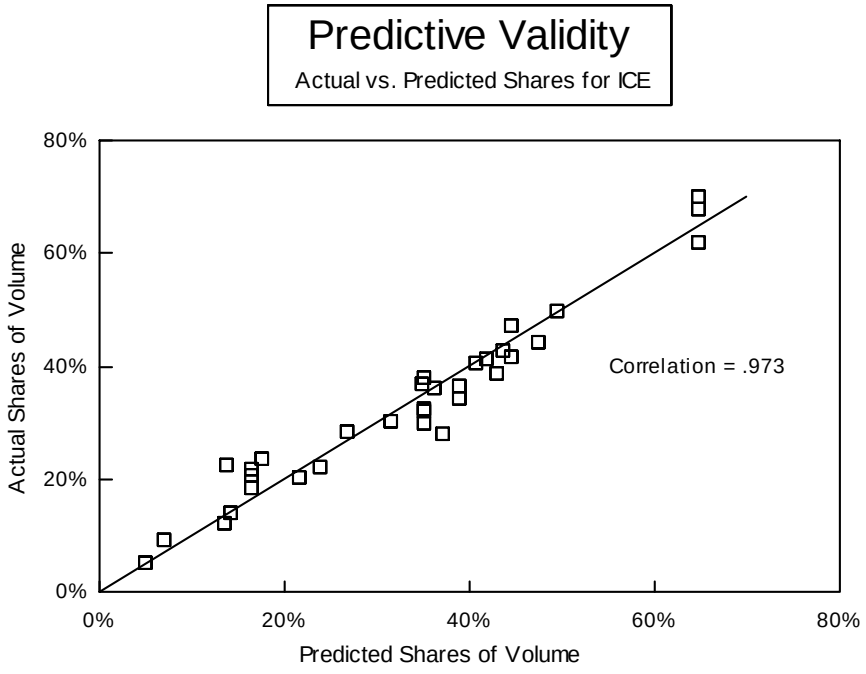
The actual and predicted shares for all validation scenarios are summarized in Table 1, using two measures of fit: MAE (Mean Absolute Error) and Correlation.

Table 1
Predictive Validity

	MAE	Correlation
Logit ME	4.65%	0.905
Logit BxP	3.92%	0.951
LC ME	3.14%	0.967
LC BxP	3.31%	0.967
ICE	2.87%	0.973

ICE and LC appear to predict actual sales better than aggregate logit. A statistical test (F-test) of the differences in predictive ability suggests that the only significant differences (at or above the 95% confidence level) are for the LC and ICE models relative to Logit ME. Ours is not the first study to conclude that capturing heterogeneity improves predictions. Studies using synthetic data (Johnson, 1997a), respondent data with holdout choices (Johnson, 1997a; Huber and Orme, 1999), and real sales data (Natter *et al.,* 1998) have also demonstrated that recognizing heterogeneity improves results. Plotting the best model (ICE) versus real market shares shows a remarkable correlation of .973 between predicted and actual shares (we've added a 45-degree line to represent the line of perfect prediction).

Figure 1



We were delighted with how accurately the conjoint simulations predicted the shares, especially in light of our relatively modest sample sizes. For the majority of the observations (categories I and III), no adjustments (scale or external effects) were made. For category II, shares for the premium and discount brands were modeled separately to adjust for a significant difference in the amount of shelf space given to the premium versus discount brands.

For our study, adding interaction terms to aggregate logit improved the predictions, but it still fell short of the simpler main-effects LC and ICE models (see Table 1). We didn't investigate additional terms to account for similarity effects and cross-elasticities, but speculate that they would have improved predictions for aggregate logit. Also note that adding interaction terms to the LC model did not improve validity, and the ICE model with no interactions terms was the most successful. This finding suggests that disaggregate approaches may somehow account for complex effects (such as interactions) with main-effects models. We'll provide more evidence on this later.

The Role of the Scale Factor

At the onset of this project, we expected we would need to adjust the scale factor (exponent) to accurately predict market shares. The scale factor controls the flatness or steepness of share estimates. The more random noise in the conjoint responses, the flatter the share predictions (and vice-versa). We were pleased (and a little surprised) that no significant tuning was justified. Our respondents evidently made choices in the CBC tasks with roughly the same degree of attention as buyers in general give to the real world purchases. This may not always be the case for other product categories and situations where conjoint is used to predict market shares.

Though there were small differences in the implied scale factor (based on the variance of the predicted shares) between the aggregate logit, LC and ICE models, one cannot argue that these account for the differences in predictive validity for our study. We experimented with tuning the exponent for the aggregate logit models, but that resulted in insignificant overall improvements in MAE and correlation. Also, computing correlations between share predictions and actual sales (as shown in Table 1) largely, but not entirely, controls for scale.

Our findings agree with other researchers who have suggested (Johnson, 1988) or demonstrated (Brice, 1997) that the Share of Preference (logit) model generally predicts actual market shares better than the more extreme First Choice model. (The First Choice model is equivalent to a Share of Preference model with a very high exponent.) Though the Share of Preference scaling worked best for our study, there may be specific instances (high involvement purchases, relatively noisy conjoint data) where the First Choice model excels.

Notes and Caveats for Predicting Sales with CBC

Why did our CBC simulations predict sales so well?

1. We exercised considerable control by interviewing respondents in the same stores that contributed the validation sales data.

2. Brands in Categories I and III were equally available on the shelves (roughly same shelf space) in all five stores. CBC interviews mimic this presentation, since each brand is equally represented (in terms of computer screen real estate) in choice tasks. Category II had very unequal representation on the shelf between the premium and discount brands, so we separately modeled shares among them.

3. Conjoint analysis reflects stable, long-term demand. None of the products we studied were new introductions, and we carefully selected validation periods by accumulating sales data across weeks in which the prices had remained constant for enough time to reasonably stabilize demand. As mentioned before, we retained 71% of the total sales data available to us.

Additionally, we speculate that the following helped:

4. Pictorial representation of the products.

5. Conditional pricing, to reflect realistic prices.

6. We chose categories that had few brands and included only the most commonly purchased package size of each.

Now that we have demonstrated that ICE and LC outperformed aggregate logit in predicting actual sales for our study, we'll spend the remainder of this paper discussing why capturing heterogeneity improves predictive validity.

IIA (Independence from Irrelevant Alternatives)

The logit model (described later) is governed by a property called IIA (Independence from Irrelevant Alternatives). The property dictates that products in a market simulation take share from other products in proportion to their respective shares. At first, this property was thought beneficial (Johnson, 1997b), but it poses problems for accurate market representations. Unless terms are specifically modeled to directly capture complex effects, aggregate level logit models ignore unequal substitution patterns between products, differential cross-elasticities, and interactions.

The "Red Bus/Blue Bus" Problem

IIA is often illustrated with the "Red Bus/Blue Bus" example. In that example, different modes of transportation are available, such as cars and red buses. The example reflects a classic line extension problem wherein the bus company decides to repaint half of its buses blue in hopes of increasing bus ridership. Although we wouldn't expect that move to significantly increase

ridership (since color is very unimportant to potential bus riders), aggregate level logit models will unrealistically inflate the net bus ridership due to the IIA property.

Our grocery store data can be used to demonstrate this principle, and how capturing heterogeneity helps resolve the IIA problem for classic line extension problems. One of the three product categories we studied was a product that came in a plastic bottle. In addition to brand and price, we included a "decoy" attribute that specified whether the bottle was round or square. The shape of the bottle was virtually unimportant to our respondents, both on average, and within all six LC segments we analyzed.

The simulation results below show the Share of Choice under main-effects aggregate level logit for five fictitious products, each at their average price:

Brand and Form	**Share of Choice**
Brand A, Square	16.7%
Brand B, Round	17.9
Brand C, Square	8.4
Brand D, Round	28.6
Brand E, Square	28.4
	100.0%

Suppose the brand manager for Brand A wanted to know if offering his product in both a square and a round bottle would significantly increase its net share. We can use the conjoint simulator to respond to that question. But depending on whether we capture heterogeneity or not, we'll get a vastly different answer. The simulation below is the same as above, but with Brand A also offered in a round bottle:

Brand and Form	**Share of Choice**
Brand A, Square	14.4%
Brand A, Round	*13.7* (Net Brand A = 14.4 + 13.7 = 28.1)
Brand B, Round	15.5
Brand C, Square	7.2
Brand D, Round	24.7
Brand E, Square	24.5
	100.0%

The aggregate level logit simulator suggests that the line extension will increase the net share to Brand A from 16.7% to 28.1%, or a 68% increase. If you were the brand manager, would you believe it?

Capturing heterogeneity with LC or ICE results in more realistic answers. Using the same simulation scenarios as above, we calculated the net increase to Brand A for the line extension under 2- through 8-group LC solutions, and with ICE.

Table 2
Line Extension Example
Relative Share Increase for Brand A

Method	**Increase**
Aggregate Logit	+68%
2-Group LC	+47%
3-Group LC	+37%
4-Group LC	+31%
5-Group LC	+24%
6-Group LC	+25%
7-Group LC	+19%
8-Group LC	+23%
ICE	+11%

The share inflation problem is reduced considerably as we recognize more heterogeneity. ICE (which fits a set of utilities for each individual) shows the least amount of share inflation—and we'll argue is probably the most realistic answer.

Why does capturing heterogeneity reduce share inflation for similar products? Consider the case of a LC solution. Utilities customized within each group of respondents fit respondent choices better than a single aggregate set of logit utilities. When logit/LC utilities better fit the data, they become larger (in absolute value). The shares of preference for products *within* each segment become more extreme because the utilities have greater variance. (This is the same familiar situation as increasing the exponent in Sawtooth Software simulators.) As utilities become more extreme, simulations begin to behave more like the First Choice model within each segment, which is immune to IIA problems. When results are averaged across segments, the overall scaling is very close to the original aggregate level logit, so the sensitivity of the market simulation model remains largely unaffected.

Cross-Elasticities

Another weakness of aggregate-level logit is its inability to account for cross-elasticities with the standard main-effects or main-effects-plus-interactions models. (One can model cross-elasticities with aggregate-level logit, but that requires including many more parameters to be estimated.)

Cross-elasticity is defined as the relative percent change in quantity demanded of brand A resulting from a percent change in price of brand B. Recall that with aggregate-level logit, IIA dictates that when a brand lowers its price, it steals share from other brands in proportion to the other brands' shares. In other words, the cross-elasticities are held constant.

We can use the line extension example from the previous section to illustrate cross-elasticity. Recall that we ended up with six total products (after Brand A was released in both the square and round bottle). What happens if Brand A, square bottle lowers its price by 20%? Aggregate logit simulations reveal the following:

Table 3
Constant Cross-Elasticities under Aggregate Main-Effects Logit

Product	Share at Avg. Price	Brand A, Square Lowers Price by 20%	Percent Change in Share
Brand A, Square	14.4	21.7	+51%
Brand A, Round	13.7	12.6	-8%
Brand B, Round	15.5	14.2	-8%
Brand C, Square	7.2	6.6	-8%
Brand D, Round	24.7	22.6	-8%
Brand E, Square	24.5	22.4	-8%

The results are consistent with IIA (constant cross-elasticities). In reality, we'd expect the square form of Brand A to take relatively more share away from the round form of the same brand than from the remaining products. ICE simulations suggest that behavior:

Table 4
ICE Cross-Elasticity Example

Product	Share at Avg. Price	Brand A, Square Lowers Price by 20%	Percent Change in Share
Brand A, Square	9.3	17.0	+83%
Brand A, Round	9.3	7.0	-25%
Brand B, Round	15.6	13.6	-13%
Brand C, Square	6.4	5.4	-16%
Brand D, Round	33.1	31.2	-6%
Brand E, Square	26.3	25.8	-2%

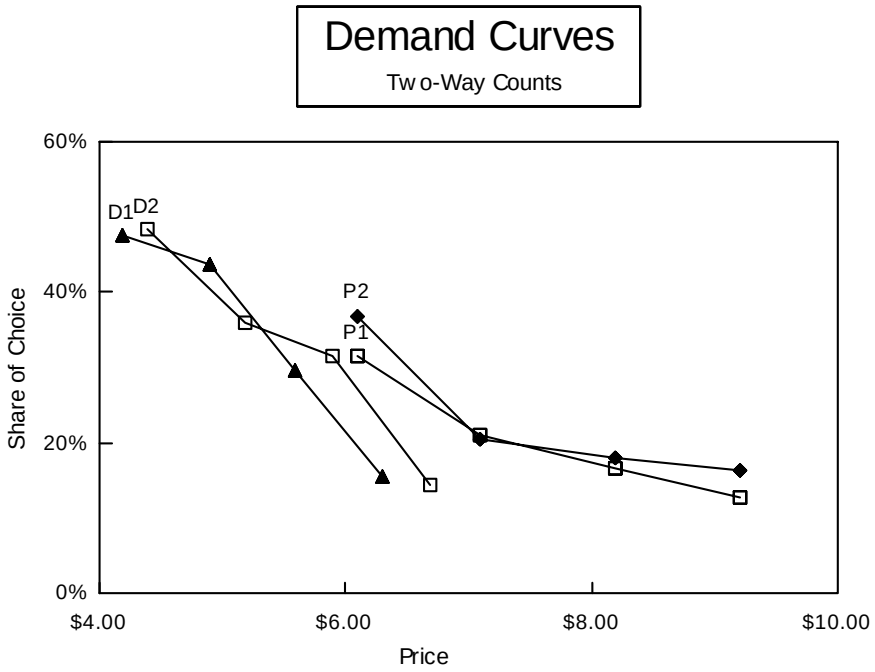
If we account for respondent heterogeneity (with LC or ICE), some degree of differential cross-elasticity can be captured and modeled using only main-effects models.

Interactions

One of the celebrated benefits of aggregate-level logit is the ability to model interactions, such as those between brand and price. However, if interactions result from differences in preference between segments or individuals, these can also be captured by recognizing heterogeneity with LC or ICE *without having to include interaction terms*.

The following graph displays two-way Count probabilities from CBC, representing the probability of each brand being chosen at each of its price points. Brands P1 and P2 are two premium brands. D1 and D2 are two discounted brands.

Figure 2



Each brand was shown at customized price points, to reflect its realistic price range. The data suggest an interaction between brand and price. The premium brands' shares do not appear to be as elastic as the discount brands. A log-log regression of share of choice on price reflects the following elasticities, and confirms that observation:

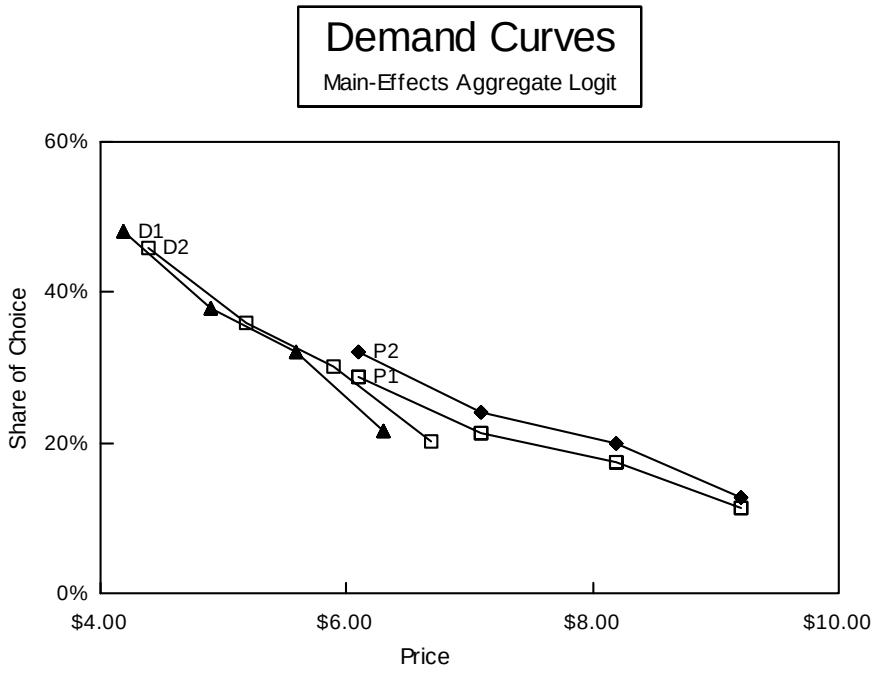
Table 5
Computed Elasticities from
Two-Way CBC Counts

	Elasticity
P1	-2.15
P2	-1.89
D1	-2.72
D2	-2.65

We can also generate demand curves based on the same data using conjoint simulators. To do so, we conduct sensitivity analysis. For the example above, four products are entered in the market simulator. To determine the demand curve for a brand, we compute its share of choice at each price point, holding the other brands constant at their average prices.

Under main-effects only simulators, a single set of price utilities is calculated to reflect the average impact of price on choice, everything else (including brand) held constant. Computing demand curves for our example (which appears to exhibit an interaction between brand and price) using aggregate logit would be improper:

Figure 3



The interaction effect is lost. Table 6 displays the elasticities calculated from the aggregate logit main-effects only model next to those from Counts.

Table 6
Elasticities

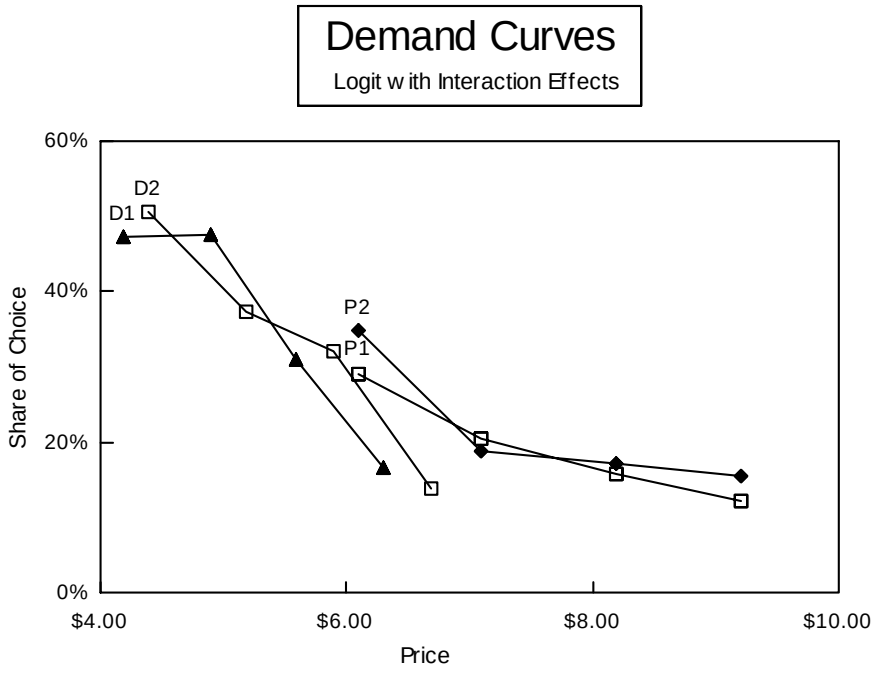
	Two-Way Counts	Aggregate Logit ME
P1	-2.15	-2.16
P2	-1.89	-2.11
D1	-2.72	-1.86
D2	-2.65	-1.88

The elasticities are nearly constant under main-effects aggregate logit—the only discrepancies due to the average difference in height of the demand curves. (Given parallel demand curves,

the higher share brands have lower elasticities, because a loss of each share point reflects a smaller percentage of the base share than for smaller share brands.)

Adding interaction terms between brand and price to the aggregate level model results in a much better fit to the underlying Counts data:

Figure 4



And the elasticities closely match the counts data:

Table 7
Elasticities

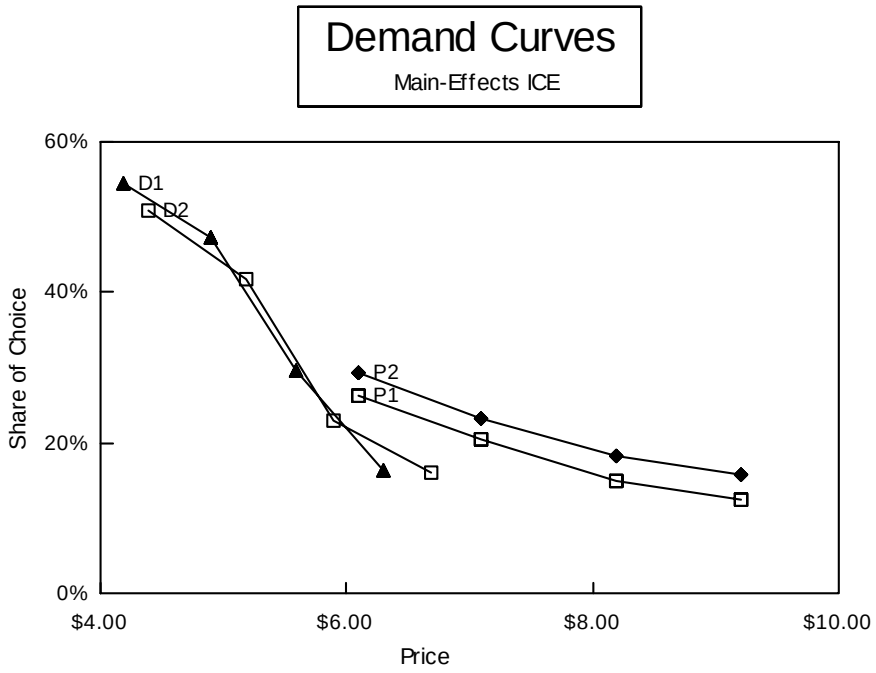
	Two-Way Counts	Aggregate Logit: Main-Effects Only	Aggregate Logit with Interactions
P1	-2.15	-2.16	-2.07
P2	-1.89	-2.11	-1.85
D1	-2.72	-1.86	-2.57
D2	-2.65	-1.88	-2.85

We've seen that aggregate logit requires interaction terms to adequately model the relationship between brand and price. The main-effects only model failed to recognize the difference in

elasticity between brands under aggregate logit, but how will the same model perform if we recognize heterogeneity?

Demand curves for the four brands under ICE (main-effects only) are as follows:

Figure 5



Even though we used main-effects only (did not include an interaction term for brand x price) the demand curves reflect that the premium brands are less price sensitive than the discount brands. Note that the price curves developed under main-effects ICE are generally smoother than the counts data and the logit with interactions model. In our opinion, ICE seems to cut through a lot of the noise (we don't have a particularly large data set) and do a fairly good job in reflecting the underlying relationships suggested by the Counts data. Table 8 reflects elasticities for all the models presented thus far.

Table 8
Elasticities

	Two-Way Counts	Aggregate Logit: Main-Effects Only	Aggregate Logit with Interactions	ICE: Main-Effects Only
P1	-2.15	-2.16	-2.07	-1.86
P2	-1.89	-2.11	-1.85	-1.53
D1	-2.72	-1.86	-2.57	-2.97
D2	-2.65	-1.88	-2.85	-2.86

How is it that ICE can reflect differential price elasticities using only a main-effects model? It is because the same respondents who strongly prefer the premium brands are also less price sensitive. Respondents who prefer (or are willing to settle for) the discount brands are generally more price sensitive. Those individuals who choose the premium brands in simulations will be less likely to change to the discount brands due to price movements, and vice-versa. It is important to recognize that *any* main-effects conjoint model that captures heterogeneity (traditional conjoint, ACA) can detect interactions when conducting sensitivity simulations.

We are not necessarily advocating modeling demand curves with main-effect disaggregate models rather than counts tables or aggregate models that directly model the interaction between brand and price, but pointing out that such models may often do an adequate job of reflecting more complex relationships without the additional terms added.

Summary and Concluding Remarks

Our data show that, under favorable conditions, CBC can accurately predict market shares for packaged goods at the grocery store. This extremely encouraging result calls for replication and verification. Our data also demonstrate that recognizing heterogeneity can improve results.

Aggregate-level logit has been faulted for its IIA properties. Specifically, it can fail when products with differing degrees of similarity are included in simulations. Corrections for product similarity such as Huber, Orme and Miller's RFC model (Huber, Orme and Miller, 1999) can improve aggregate level logit simulations, but it is best to begin with an underlying model that is less susceptible to the "Red Bus/Blue Bus" problem.

Many failings of IIA can be avoided by adding complex terms to aggregate logit models (i.e. interaction terms, cross-elasticities, availability terms). These models can become very complex and risk becoming over-fitted (too many terms relative to observations), fitting a good deal of noise along with true effects. It puts a heavy burden on the analyst to choose the right

combination of complex terms to maximize predictive validity for aggregate logit models. But before the advent of disaggregate choice modeling techniques, more complex specifications were often needed to model the marketplace adequately.

Using LC and especially ICE can reflect complex effects (differential substitution, cross-effects and interactions) and achieve very accurate predictions using parsimonious main-effects models if such effects can be largely accounted for by differences in preference among underlying segments or individuals. Our research adds to a growing body of evidence that suggests this is often the case.

References

Brice, Roger (1997), "Conjoint Analysis A Review of Conjoint Paradigms and Discussion of the Outstanding Design Issues," *Marketing and Research Today*, November, 260-66.

Clancy, Kevin J. and Robert S. Shulman (1994), Marketing Myths that are Killing Business, New York: McGraw Hill, Inc., 201.

Huber, Joel, Bryan Orme and Richard Miller (1999), "Dealing with Product Similarity in Conjoint Simulations," *Sawtooth Software Conference Proceedings*.

Johnson, Richard M. (1988), "Comparison of Conjoint Choice Simulators–Comment," *Sawtooth Software Conference Proceedings*, 105-108.

Johnson, Richard M. (1997a), "Individual Utilities from Choice Data: A New Method," *Sawtooth Software Conference Proceedings*, 191-208.

Johnson, Richard M. (1997b), "Getting the Most from CBC—Part 2," *Sawtooth Solutions*, Spring, 2-3.

Natter, Martin, Markus Feurstein and Leonhard Kehl (1998), "External Validity of Segmentation Based CBC-Analysis," Marketing Science Conference, Fountainbleu, France.