A Marketing Mix Model Developed From Single Source Data: A Semiparametric Approach

by

Makoto Abe

S.B. Electrical Engineering and Computer Science, M. I. T. (1984)

S.M. Electrical Engineering and Computer Science, M. I. T. (1984)

Submitted to the Department of Physics
in partial fulfillment of the requirements for the Degree of

Doctor of Philosophy in Operations Research

at the

Massachusetts Institute of Technology

July 1991

Signature of Author _____
Department of Physics, Operations Research

Certified by _____
John D. C. Little, Institute Professor, Thesis Advisor

Accepted by _____
George F. Koster, Chairman of Graduate Committee

A MARKETING MIX MODEL DEVELOPED FROM SINGLE SOURCE DATA: A SEMIPARAMETRIC APPROACH

by

Makoto Abe

Submitted to the Department of Physics
in partial fulfillment of the requirements
for the Degree of Doctor of Science in Operations Research

ABSTRACT

Recent advances in information technology have made available large single source databases which contain household purchase and shopping trip records collected by UPC scanners and advertising exposures by TV meters. Such databases permit analyses on a household level and have opened up a whole new direction in marketing. The issues of interest cover a wide range from brand choice, purchase quantities, and interpurchase timing to behavioral theories of price, advertising, and promotion response, as well as repeat purchasing. The theme of this dissertation throughout is how to obtain the most use out of such enormous amounts of data.

Taking advantage of the data size, Part I approaches modeling from a different angle by discarding the parametric statistical models and using empirical joint densities of relevant variables. The method of nonparametric density estimation (NDE) is compared with multinomial logit model (MNL) --- a popular parametric method in consumer brand choice. While the empirical results of NDE show promise, the method requires an enormous amount of data, even beyond the scope of scanner data. This sets practical limitations on the approach.

This conclusion leads to Part II, where a middle approach between parametric and nonparametric methods is pursued. A semiparametric utility residual method (URM) is proposed that retains the assumption of stochastic utility maximization and the extreme value distribution of MNL while relaxing the linear utility function by using additive one dimensional nonparametric functions of explanatory variables. Part II conducts an extensive simulation study to investigate the operational characteristics of URM, and then applies the method to two actual scanner databases to illustrate its power.

Part III focuses on category purchase incidence in order to pursue household level analyses of sales in addition to brand choice and share as considered in Part I and II. The model is based on a nested logit driven by shopping trips, and URM is employed for graphical diagnostics to infer appropriate parametric utility transformations. The URM procedure is found to be quite useful in identifying influential points, outliers, and heterogeneous segments.

Finally, Part IV adds a Poisson advertising exposure model to the nested logit marketing mix model calibrated in Part III. The exposure model computes a household advertising stock variable that is an input of the marketing mix model. This is done by converting GRPs by week and daypart to household adstocks, taking into account household media habits. The combined model permits a simulation of various ad scenarios to evaluate their sales and share implications.

Thesis Supervisor: Professor John D. C. Little

Title: Institute Professor

ACKNOWLEDGEMENTS

This thesis, in some way, summarizes my enjoyable 12 years at MIT with help from many people. First and foremost, I am indebted to my supervisor, John D. C. Little, for training me as a researcher in Marketing. There is not simply sufficient space to describe how grateful I am. He introduced Marketing to me as an Operations Research student four years ago and taught me everything to know about it from A to N ever since. Now I must learn from O to Z on my own. I must apologize for referring to many of his articles in Part IV, which are solely based on my own judgement.

I own John Hauser for his sharp advices on this thesis as well as other papers. His appropriate comments always seem to hit the crucial points of the forefront of Marketing. I also wish to thank to the other member of the committee, Moshe Ben-Akiva and Tom Stoker for their time and helpful comments. Whenever there was a committee meeting, Moshe had to walk all the way from his office in the main building.

I appreciate John Tarsa of Ocean Spray Cranberries, Inc. for making the single source data available and providing me with valuable managerial insights at many meetings with Professor Little. I would also like to thank people at Information Resources Inc., especially Bob Brooks and Doug Honnold for preparing data.

Not to be forgotten is Wujin Chu. Although not directly involved in this thesis, Wujin has been always accessible for discussion inside as well as outside of academics. The other faculty, Birger Wernerfelt, Bill Qualls, and France LeClerc have been very supportive whenever I have silly questions. Patty Shaughnessy was the person responsible for creating the friendly atmosphere of Marketing group.

CONTENTS

A Nonparametric Density Estimation Method for Brand Choice Using Scanner Data

Makoto Abe

Operations Research Center
M.I.T.
Cambridge, MA 02139 USA

M. I. T. Doctoral Dissertation, Part I

June 1991

OVERVIEW

Recent advances in scanner technology have made available large databases of individual purchase records and opened up a whole new direction in marketing science. The issues of interest cover a wide range from brand choice, purchase quantities, interpurchase timing to behavioral theories of price, advertising, and promotion response as well as repeat purchasing. In studying these, many models have been created to address specific questions to databases which contain enormous amount of information. Most of these models are parametric in nature, in other words, a specified model contains some number of unknown constants (parameters). These are then estimated from the data, various tests are made, and conclusions are drawn. A weakness in this methodology is that the underlying specification of the model may be incorrect, in which case the parameter estimates will be biased and subject to misinterpretation. For example, in discrete choice models both probit and logit make a distributional assumption about the stochastic term, and the logit further posits independence from irrelevant alternatives. In inter-purchase timing, many models assume parametric distribution functions such as gamma or lognormal. For verifying theories, an elaborate modeling scheme must often be devised in which special variables and sequences of models are created to facilitate hypothesis testings.

This paper takes an alternative path by utilizing nonparametric methods. These are appealing because they make few or, at least, fewer underlying assumptions and offer great structural flexibility. In particular, a comparison of nonparametric density estimation using a kernel method with multinomial logit for modeling consumer brand choice is investigated using IRI scanner data. The tracking results indicate that the nonparametric results are superior to the logit for this database. Then, advantages and limitations of the nonparametric density estimation in more general settings are discussed, and insight is gained about modeling philosophy.

1. INTRODUCTION

Recent advances in UPC scanner technology have made available large databases of individual purchase records. A brief calculation shows that about 2 gigabyte (2,000 megabyte) of data is generated every week across the nation. This enormous amount of data has caused a data glut situation as noted in August 28, 1989 issue of Business Week (p57).

> *In the meantime, they* [the companies] *are trying to figure out how to make better use of the reams of information they already have. ... "The information preceded clients' ability to handle it," says SAMI head Steven A. Wilson. "They can't absorb the amount of information we're imposing on them." Adds Brian M. Shea, Ore-Ida's marketing research manager.*

From a research point of view, however, there is a positive side in this data glut. It has created an unprecedented opportunity for new approaches in modeling, which is a key determinant of the performance of marketing decision support systems. In most experimental disciplines, the accuracy of the study is limited by the amount of data used, and more samples are sought if time and cost allow. Most techniques in experimental design such as fractional factorial designs stem from the common goal of extracting the maximal information from a given number of data points. By contrast, our situation is rare in that there exists more data than we can easily handle. Therefore, some non-traditional, if not radical, approaches may be in order.

For example, when a traditional econometric discrete choice model, multinomial logit (MNL), is applied to a panel data obtained from the scanner, it is common to observe t-values for some coefficients such as loyalty to be as high as 20 or more. (Guadagni & Little 1983) This is quite natural considering that a dozen or so parameters are estimated from thousands of observations. Obtaining such high t-values is great news as long as the underlying model is correct. But in many cases, it is not an easy task to specify the model appropriately, and that's why so many specification tests exist. For instance, in the OLS regression, an entire textbook can be devoted to such tests as residual analyses, optimal transformations, and tests for multicollinearity. MNL is no exception.

With abundant data, perhaps we can afford to let the data do more of the work of specifying the model structure, even at some loss of efficiency, instead of having a data analyst iterate the process of postulating a model and conducting specification tests on it. That's where nonparametric methods come in. They are based on fewer assumptions and possess much more

structural freedom so that information contained in the data is preserved without the bias or distortion that can arise from the model structure itself.

Figure 1 demonstrates the power of nonparametric methods. It is a nonparametric regression fitted to points which are generated by a quadratic underlying model with normal disturbances. The solid line is a fit by nonparametric regression called moving ellipsoid method (MEM) (Abe 1991) while the dashed line is the ordinary least squares (OLS). As expected, the OLS is a straight line which hardly resembles the true model. Of course, if the analyst knows that the underlying model is quadratic ex-ante, he or she would certainly add a quadratic term. While one could visually conjecture this fact from the plot in Figure 1, such inspections become increasingly difficult in higher dimensions. Eventually one must rely on various statistical tests and follow a trial and error iterative process to specify the correct model. In contrast, the nonparametric method is much more automatic. It is necessary to specify smoothing constants but some latitude is permitted in their values. Thereafter, once the data is entered, pressing the return key is all it takes to obtain the result.

The current research investigates the feasibility of one particular nonparametric method, nonparametric density estimation, in the context of brand choice modeling. The method is compared against a popular parametric counterpart, multinomial logit model (MNL), on household-level scanner data, and its advantages and disadvantages are evaluated systematically. Section 2 illustrates the basic theory of the nonparametric density estimation using a kernel method in discrete choice. Then in Section 3, application to scanner data for modeling consumer choice is demonstrated and a comparison with MNL is made from various angles. In Section 4, analyses of the results lead to its pros and cons over MNL under general settings, followed by a discussion of modeling philosophy and concluding remarks in Section 5.

2. A NONPARAMETRIC DENSITY ESTIMATION METHOD

2.1 Concept and Basics

A wide range of marketing studies lies on obtaining a conditional expectation of a response variable y given a set of explanatory variables, $\mathbf{x}$, $E(y|\mathbf{x})$. For example, y could be a brand choice, interpurchase time, or quantity purchased. A vector $\mathbf{x}$ could include marketing mix variables (e.g. price, promotions), product attributes, and buyer characteristics (e.g. income,

Figure 1

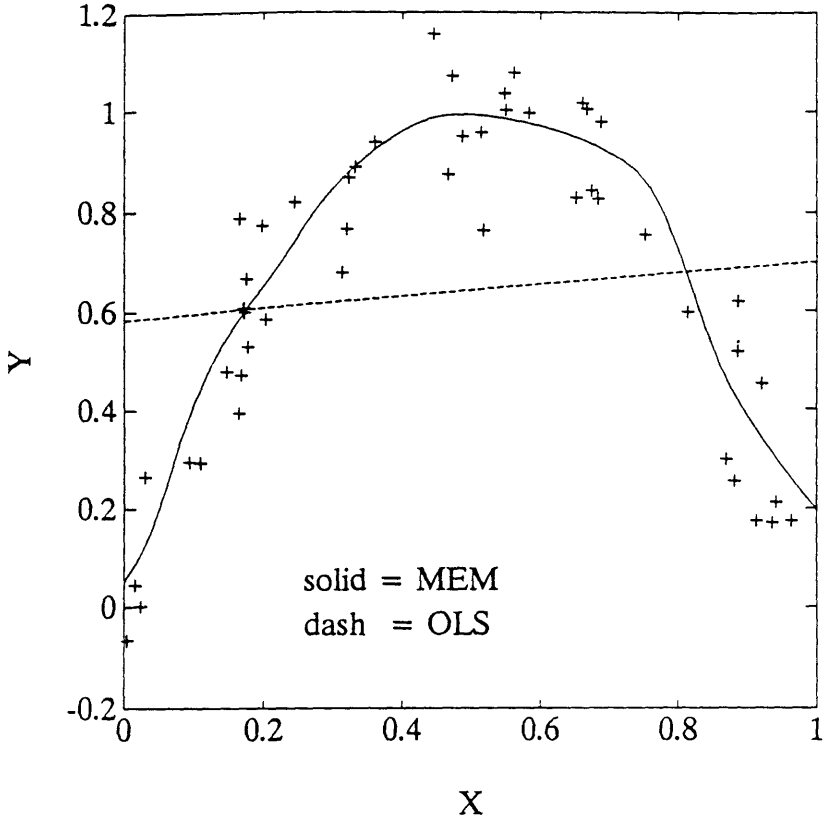


Figure 1: Example of a Nonparametric Regression

Although visual inspection can prevent mistakes like the OLS line in this two dimensional data, in the multivariate case, parametric methods may make serious errors without the researcher realizing it. Nonparametric methods should be much better.

family size), and may be a mixture of continuous, and nominal and ordinal discrete variables. For this class of general regression analyses, a parametric method, which assumes a certain parametric functional form and error distribution for estimating parameters like OLS and MNL, has been traditionally used. An alternative approach which does not involve such assumptions is called nonparametric regression.

When response variable y is continuous, kernel regression (Watson 1964, Nadaraya 1970) which is introduced in Marketing by Rust (1988) is one of the most popular nonparametric regression techniques. A class of splines is another well-known method. The current research, on the other hand, focuses on brand choice where the response variable is nominal discrete. In this case, more intuitive interpretation of the conditional expectation $E(y \mid x)$ or equivalently $P(y \mid x)$, is in terms of $f(x \mid y)$ --- a conditional probability density function of attribute variables x given that brand y is chosen --- as follows.

$$(1) \qquad E(y \mid x) = P(y \mid x) = \frac{f(x,y)}{f(x)} = \frac{f(x \mid y)\, P(y)}{\sum_{y} f(x \mid y)\, P(y)}$$

The regression is solved by estimating $f(x \mid y)$ nonparametrically for each brand y. We shall refer to this process as Nonparametric Density Estimation (NDE) method.

In the following discussion, the conditioning of y in $f(x \mid y)$ is suppressed for clarity on the understanding that the density function is estimated only from a subset of sample with observed choices of brand y. Some of the common nonparametric estimators for a density function, f(x), are histogram, moving average, the kernel estimators, the nearest neighbor method, orthogonal series methods, (all of these in Silverman 1986), the maximum penalized likelihood method (Good & Gaskins 1980), and spline methods (Wegman & Wright 1983). Many of these estimators, however, become rapidly complex in higher dimensions. For instance, popular splines are largely applied in one dimension and the formulation for even a two dimensional case is rather complex (Wegman & Wright 1983). In this study, the kernel method is chosen, since it remains analytically and computationally tractable in higher dimensions which are usually the case here.

The kernel estimator for the probability density function (PDF), f(x), is

(2) $$\hat{f}(x) = \frac{1}{N h^d} \sum_{i=1}^{N} k\left(\frac{x - x_i}{h}\right)$$

N is the number of observations, x_i is the i-th observation of the vector of explanatory variables x whose dimension is d. h is a smoothing constant which controls the trade off between roughness and fit of the estimator. $k(\bullet)$ is a so called kernel function, and typically but not necessarily possesses the following properties:

[a] It is a decreasing function of the argument in the appropriate metric.
[b] It is smooth and has derivatives of high order.
[c] It is non-negative and integrates to 1.

A symmetric probability density function whose density is concentrated around 0 will satisfy all of the above and it is most often used in practice. By [a] above, the form of the kernel function differs whether the argument is continuous, ordered discrete, or nominal discrete, depending on the choice of the metric. For instance, a Gaussian function is a popular choice in continuous cases, while a geometric weighting function, $k(x-x_i) = \text{constant} \times \mu^{|x - x_i|}$ (where $\mu<1$), is often used in ordered discrete cases.

The parameters h and μ are called smoothing constants, and determine how fast the value of $k(\bullet)$ falls as the argument increases. They affect smoothness of the resulting density function and balance the trade off between bias of the estimator and its variance. It is generally known that the constructed density function is quite sensitive to the value of the smoothing constant, but much less sensitive to the shape of the kernel function (Ullah 1988). The next section discusses the smoothing constant in detail and introduces some existing methods for determining its value.

When the density function is estimated only from a subset of sample with observed choice of brand y in (2), it becomes $f(x \mid y)$. Substituting this in (1) provides our NDE model of choice probability $P(y \mid x)$. Note that unlike MNL, the probability is derived without any reference to utility maximization, a linear-in-parameter utility function, or a doubly exponential stochastic component.

It turns out that equation (1) can be also obtained by direct application of kernel regression when the response variable is 0/1 binary discrete as shown below. A general expression for kernel regression of y on x is (whether y is continuous or discrete)

$$E(y \mid x) = \frac{\sum_i y_i \, k\left(\frac{x-x_i}{h}\right)}{\sum_i k\left(\frac{x-x_i}{h}\right)}$$

where subscript i is an index for an observation.

If **y** is a vector of J elements (each corresponding to one of J alternatives) with j-th element y_j being 1 if alternative j is chosen and 0 otherwise, then $E(y_j \mid x) = P(y_j \mid x)$, and the above regression becomes

$$\begin{aligned} E(y_j \mid x) \equiv P(y_j \mid x) &= \frac{\sum_{i=1}^{N} y_{ij} \, k\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{N} k\left(\frac{x-x_i}{h}\right)} \\ &= \frac{\sum_{i \in \text{choose brand } j}^{N_j} k\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{N} k\left(\frac{x-x_i}{h}\right)} \\ &= \frac{\frac{N_j}{N} \frac{1}{N_j} \sum_{i \in \text{choose brand } j}^{N_j} k\left(\frac{x-x_i}{h}\right)}{\frac{1}{N} \sum_{i=1}^{N} k\left(\frac{x-x_i}{h}\right)} \\ &= \frac{P(y_j) \, f(x \mid y_j)}{f(x)} \end{aligned}$$

Hence, in the case of discrete response variable y_j, the nonparametric kernel regression reduces to the nonparametric density estimation of **x** conditional on choosing brand j.

2.2 Smoothing Constant

The value of h must be determined based on some criteria, ideally, a function of the difference between the unknown underlying density function, f(x), and its estimate, $\hat{f}(x)$. A common

discrepancy measure, mean square error (MSE), can be decomposed into two elements, bias and variance, as,

$$MSE_x(\hat{f}) = E\left\{\hat{f}(x) - f(x)\right\}^2 = \left\{E\,\hat{f}(x) - f(x)\right\}^2 + E\left\{\hat{f}(x) - E(\hat{f}(x)\right\}^2 = \text{bias}\,(x)^2 + \text{var}\,\hat{f}(x)$$

If we limit the kernel k(t) to a radially symmetric d-dimensional PDF such that $\int k(t)dt=1$, $\int tk(t)dt=0$, and $\int t_1^2\, k(t)dt=\sigma_k^2$, then the asymptotic values of bias and variance are approximately,

$$\text{bias}\,(x) \approx \frac{1}{2} h^2 \sigma_k^2 \nabla^2 f(x)$$

$$\text{var}\,\hat{f}(x) \approx n^{-1} h^{-d} \beta\, f(x), \qquad \text{where} \quad \beta = \int k(t)^2\, dt$$

These results can be obtained by applying the Taylor series expansion (Silverman 1986). Several important observations can be made here. First, for a fixed h, the bias is independent of n while the variance is inversely proportional to n. Of course, both are indirectly related to the sample size through h which should depend on n for consistent estimation. Second, reduction of h decreases the bias but increases the variance at the same time. Thus, h balances the trade off between the magnitudes of the bias and variance. Now, the optimal h which minimizes the global discrepancy measure, mean integrated square error (MISE, L_2 criterion) defined as

(3) $$MISE(\hat{f}) \approx E\int \left\{\hat{f}(x) - f(x)\right\}^2 dx = \frac{1}{4} h^4 \sigma_k^4 \int \left\{\nabla^2 f(x)\right\}^2 dx + n^{-1} h^{-d} \beta$$

can be found to be

(4) $$h_{opt} \approx \left[n \frac{\sigma_k^4}{d\,\beta} \int (\nabla^2 f)^2\, dx \right]^{-\frac{1}{d+4}}$$

by the first order condition. Note that it depends on the form of the kernel via β and σ_K^2 and on the second derivative of the unknown true density function. The rougher the function is, the smaller the optimal h is. It is known that the MISE is rather insensitive to the functional form of the kernel. Also, the smoothing constant, h, converges to 0 rather slowly ($\sim n^{-1/(d+4)}$) as the sample size increases.

Because h depends on the unknown density function, there are several methods for selecting its value.

[1] Subjective approach: Human judgement from pictorial plots, assuming the true PDF as some convenient parametric function, and the test graph method (Silverman 1986).

[2] Cross validation (least square, likelihood): Carrying out estimation by removing one sample and comparing its prediction with the actual value.

[3] Internal estimation of the second derivative of the density: Iteratively adjust h so that the consistency is obtained for the value of $\int(\nabla^2 f)^2$ which is used in h to start with and its estimation, $\int(\nabla^2 \hat{f})^2$ (Silverman 1986).

Automatic methods for finding the best h such as cross validation are a current focus in econometric research and a good survey can be found in Marron (1989).

The kernel estimator is shown to be consistent under mild regularity conditions. Since there exists an enormous literature on the proofs of the asymptotic properties of the density estimators (Prakasa Rao 1983, Devroye & Gyorfi 1985), only a few relevant ones are briefly pointed out in the appendix.

3. APPLICATION

To demonstrate the feasibility of the nonparametric density estimation (NDE) method and evaluate its performance, a comparison is made on a brand choice process with one of the most popular parametric discrete choice model in Marketing, multinomial logit (MNL).

The subset database contains store and panel data on three national aseptic fruit drink brands (each pack of three single serving paper cartons), Hi-C (brs 1), KoolAid (brs 2), and Ssips (brs 3)

during weeks of 12-29-86 through 2-6-89 (111 weeks). There are 3221 total purchases made by 143 panelists whose purchases are greater than or equal to 10 throughout this period. For simplicity, it is assumed that only one package is bought at each incident since multiple unit purchases were rare.[1] The 1988 purchases made during the first 71 weeks are used for calibration of the models and the remaining 1233 purchases in the last 40 weeks (5-9-88~2-6-89) are saved as a holdout for assessing the predictive ability of the models. Vital statistics for each brand, share, average price, fractions of promotional activities among the purchases, are found in Table 1.

Table 1: Statistics of the Database

brand	share	average price ($)	feature	display
1: Hi-C	29.8%	0.790	32.0%	12.0%
2: KoolAid	22.2%	0.894	11.5%	5.1%
3: Ssips	48.0%	0.689	9.7%	0.2%

3.2 Multinomial Logit Model

Table 2 shows the result of three nested models, M1 through M3, by MNL. As shown below, LOYALTY is defined in the same way as Guadagni and Little (1983) with the decay constant α set to 0.8.

$$LOYALTY_j(t) = \alpha \times LOYALTY_j(t-1) + (1-\alpha) \times d_j(t-1)$$

where $LOYALTY_j(t)$ = loyalty of brs j at t-th purchase incidence

$$d_j(t) = \begin{cases} 1 & \text{if brs j is purchased at t-th purchase occasion} \\ 0 & \text{otherwise} \end{cases}$$

α = decay constant

PRICE is a shelf price in dollars, and FEATURE and DISPLAY are 0/1 binary indicators of the promotional activities. ASC2 and ASC3 are alternative specific constant for brs 2 and brs 3 respectively. Since it was necessary to limit the number of continuous variables in NDE due to the memory space, the less significant PRICE variable was dropped and model M2 was selected for further comparison.

[1] When an n-unit purchase is encountered, it is decomposed into n single unit purchases.

Table 2: Result of Multinomial Logit

variable	M1	M2	M3
LOYALTY	3.342	3.406	3.393
	(32.79)	(32.06)	(31.844)
FEATURE	------	0.781	0.717
	(----)	(7.126)	(6.438)
DISPLAY	------	1.094	1.019
	(----)	(6.239)	(5.781)
PRICE	------	------	-2.836
	(----)	(----)	(-3.570)
ASC2	-0.129	0.094	0.385
	(-1.613)	(1.099)	(3.244)
ASC3	0.025	0.365	0.079
	(0.314)	(4.170)	(0.666)
loglikelihood: L(β)	**-1126.7**	**-1067.1**	**-1060.7**
ρ^2	**0.4841**	**0.5114**	**0.5144**
$\bar{\rho}^2$	**0.4827**	**0.5091**	**0.5116**

L(0) = -2184.0 (loglikelihood of equal probabilities)

$$\rho^2 = 1 - \frac{L(\beta)}{L(0)}, \qquad \bar{\rho}^2 = 1 - \frac{L(\beta) - k}{L(0)}, \qquad \text{where k is the number of parameters}$$

3.3 A Nonparametric Density Estimation Model

In NDE, a multidimensional joint PDF, f(x | y), of the same set of explanatory variables as MNL, i.e. loyalties of the first two brands, features, and displays of all brands, was constructed for each brand y. It is not necessary to consider all three loyalties since they sum up to 1. These continuous values are approximated by ordered categories of 30 to allow for numerical computation. To summarize, each joint PDF, f(x | y), corresponding to the respective brand y is 8 dimensions with the following variables.

LOYALTY1	30 point representation of ordered categorical,	0 ~ 1 in interval
LOYALTY2	30 point representation of ordered categorical,	0 ~ 1 in interval
FEATURE1	binary categorical	
FEATURE2	binary categorical	

FEATURE3	binary categorical
DISPLAY1	binary categorical
DISPLAY2	binary categorical
DISPLAY3	binary categorical

The number of cells in each PDF is $30^2 \cdot 2^6 = 57{,}600$. The kernel functions used for the ordered categorical variables are discrete approximations of the Gaussian. The smoothing constant for loyalty, h, is set to 0.1 since it has produced the best goodness-of-fit after some study as shown in Figure 5 and resulted in the most reasonable marginal probability of choice v.s. its own loyalty plot in Figure 8. This value is of the same order as the standard deviation of the variables since h in equation (2) represents a standard deviation of the Gaussian. Furthermore, assuming that the underlying density function is normally distributed, the optimal value, h_{opt}, expressed in equation (4) indicates the value to be between 0.05 and 0.5. For the binary discrete variables, features and displays, the smoothing of the geometric weighting function was introduced to avoid zero density cells. The value, μ=0.99 was chosen close to 1 to avoid any unwanted effects, and a comparison with smaller μ's indicates that it produces the best goodness-of-fit as shown in Figure 6.

The actual kernel estimator for $f(x \mid y_j)$ corresponding to equation (2) is

$$
\begin{aligned}
f(x \mid y_j) = \frac{1}{N_j} \sum_{i \in \text{choose brand } j} & \frac{1}{\sqrt{2\pi h^2}} \exp\left\{ -\frac{[\text{LOYALTY1} - \text{LOYALTY1(i)}]^2}{2h^2} \right\} \\
& \times \frac{1}{\sqrt{2\pi h^2}} \exp\left\{ -\frac{[\text{LOYALTY2} - \text{LOYALTY2(i)}]^2}{2h^2} \right\} \\
& \times \left[\mu \cdot \delta_{\text{FEATURE1, FEATURE1(i)}} + (1-\mu)(1-\delta_{\text{FEATURE1, FEATURE1(i)}}) \right] \\
& \times \left[\mu \cdot \delta_{\text{FEATURE2, FEATURE2(i)}} + (1-\mu)(1-\delta_{\text{FEATURE2, FEATURE2(i)}}) \right] \\
& \times \left[\mu \cdot \delta_{\text{FEATURE3, FEATURE3(i)}} + (1-\mu)(1-\delta_{\text{FEATURE3, FEATURE3(i)}}) \right] \\
& \times \left[\mu \cdot \delta_{\text{DISPLAY1, DISPLAY1(i)}} + (1-\mu)(1-\delta_{\text{DISPLAY1, DISPLAY1(i)}}) \right] \\
& \times \left[\mu \cdot \delta_{\text{DISPLAY2, DISPLAY2(i)}} + (1-\mu)(1-\delta_{\text{DISPLAY2, DISPLAY2(i)}}) \right] \\
& \times \left[\mu \cdot \delta_{\text{DISPLAY3, DISPLAY3(i)}} + (1-\mu)(1-\delta_{\text{DISPLAY3, DISPLAY3(i)}}) \right]
\end{aligned}
$$

where N_j is the number of observations which choose brand y_j, and $\delta_{a,b}$ is a kronecker delta function.

Using above, the conditional choice probability, $P(y_j \mid x)$, in equation (1) can be written as

$$P(y_j|x) = \frac{f(x|y_j)\, N_j/N}{\sum_{y=1}^{3} f(x|y_j)\, N_j/N}$$

where N is the total number of observations.

3.4 Tracking

Figures 2 ~ 4 exhibit the predicted and actual shares of each brand with NDE and MNL by aggregating the choice probabilities and observed number of purchases respectively for each 4-week period. The vertical line separates the calibration period (12-29-86~5-2-88) and the holdout period (5-9-88~2-6-89). A visual inspection shows that NDE has a better fit than the parametric MNL in tracking the brand choice. Two measures for their goodness-of-fits, $\overline{P}$, which is a mean probability of correct choice, and R^2, are also computed in Table 3.

Table 3: Goodness-of-Fits of the Share Plots

	sample	NDE		MNL
$\overline{P}$ (mean probability of correct choice)	*calibration*	0.7309	>	0.7016
	holdout	0.6915	>	0.6885
R^2	*calibration*	0.8882	>	0.8249
	holdout	0.8112	>	0.8046

NOTE: N = 1988 for calibration sample and N = 1233 for holdout sample
h = 0.1 and $\mu = 0.99$

For the calibration sample, NDE dominates MNL in both $\overline{P}$ and R^2, while the dominance holds by smaller margin for the holdout sample. This situation might be attributed to "overfitting" in NDE. In a parametric model, an arbitrary better fit can be achieved within the calibration sample as more explanatory variables are added. In the OLS regression, this is manifested by

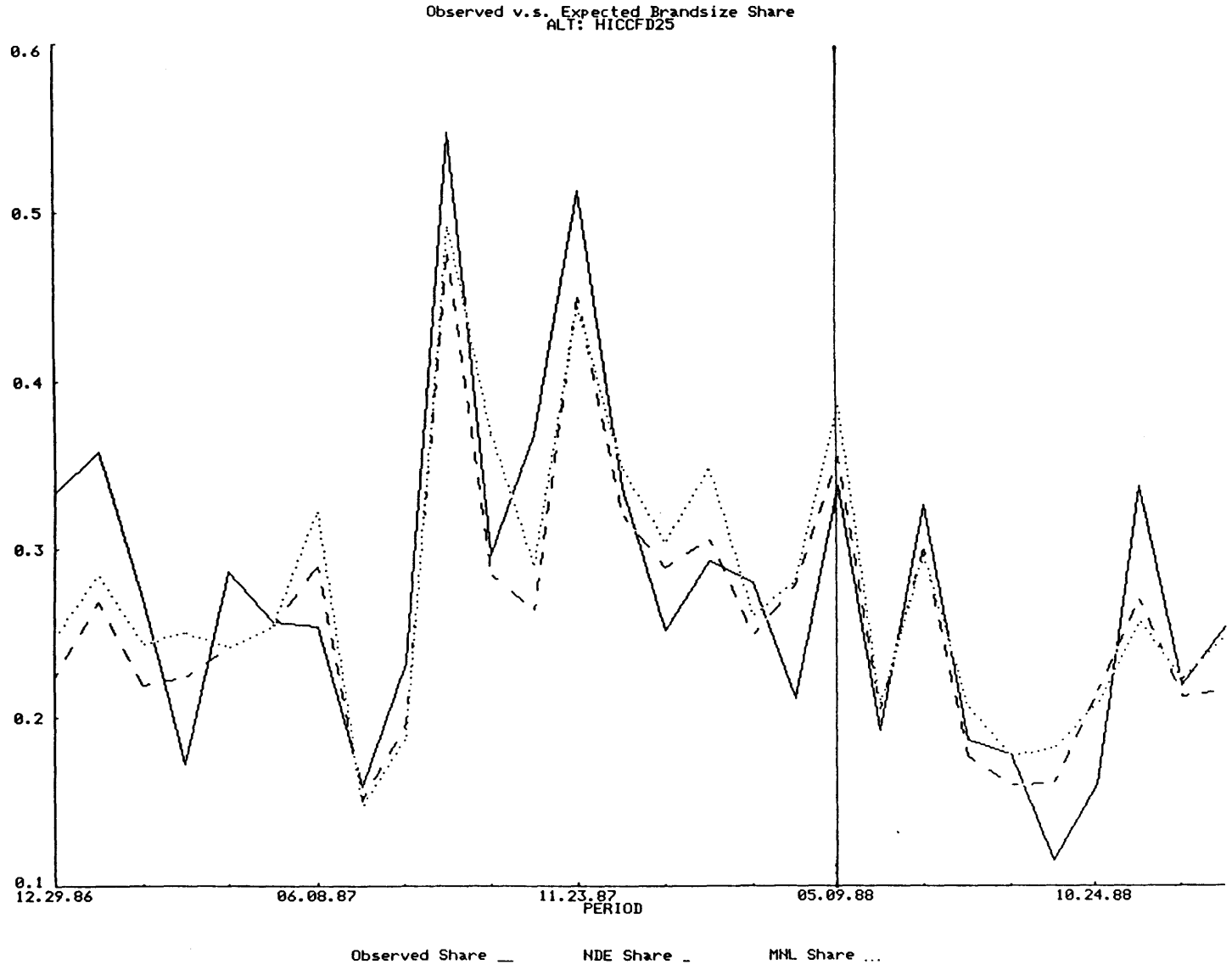


Figure 2: Time Series Share of Hi-C by NDE and MNL

Although both methods do well, close analysis shows NDE to be better.

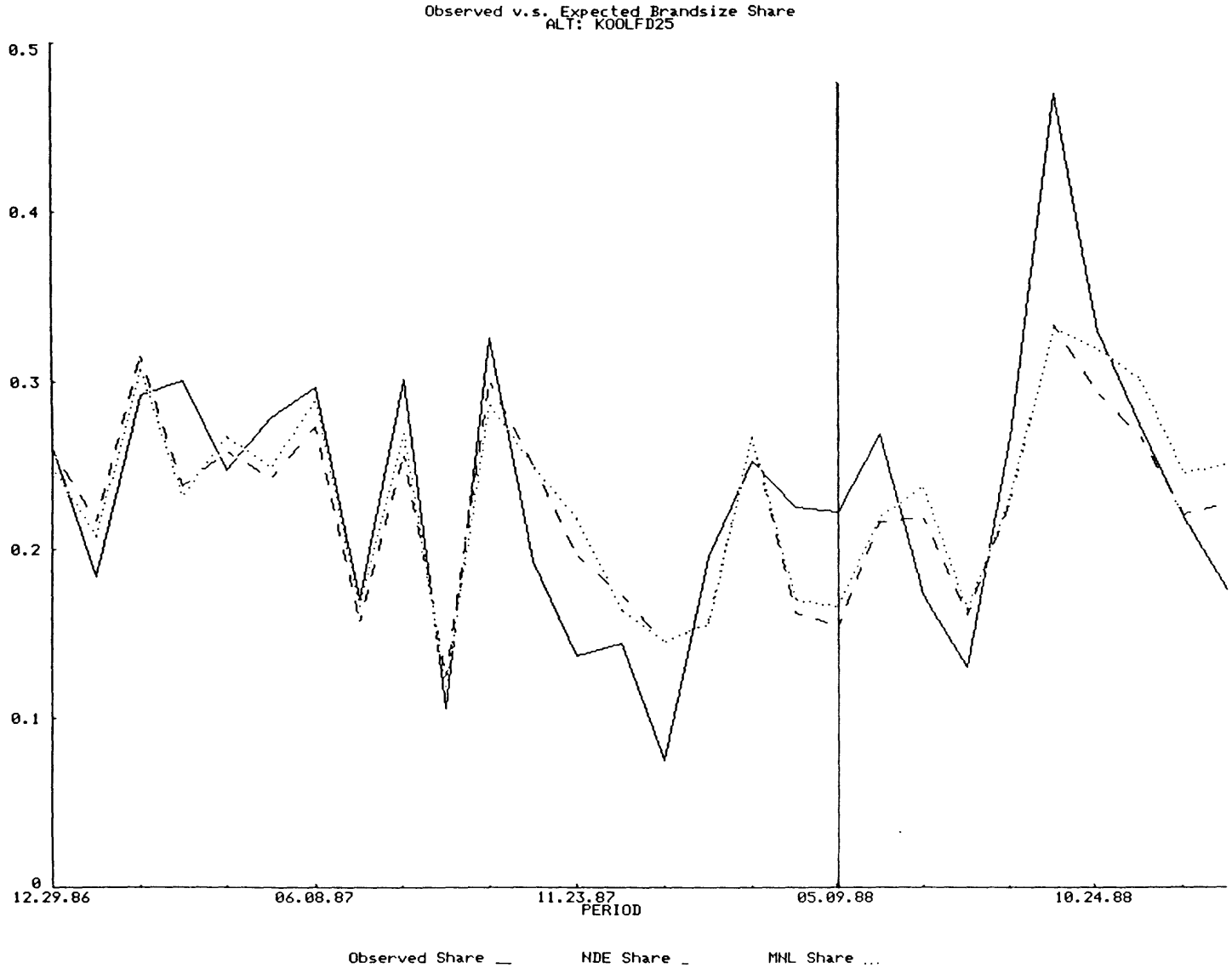


Figure 3: Time Series Share of KoolAid by NDE and MNL

Although both methods do well, close analysis shows NDE to be better.

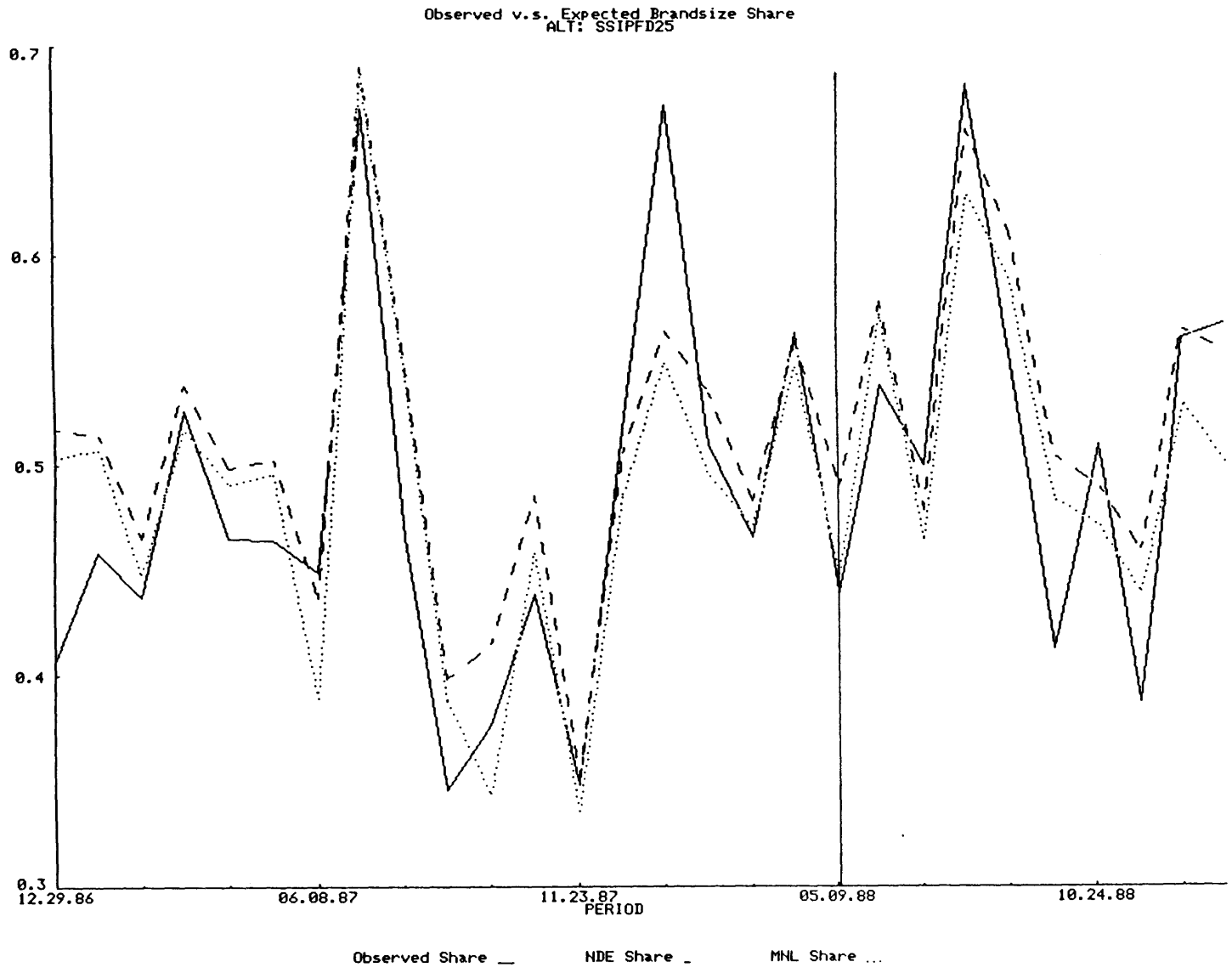


Figure 4: Time Series Share of Ssips by NDE and MNL

Although both methods do well, close analysis shows NDE to be better.

higher R^2 with more variables. But this better fit does not necessarily lead to a better prediction in the holdout sample, since the overfitted model also starts to pick up the unwanted effect of the random disturbance term which is idiosyncratic to the calibration sample. Because NDE with small h could be regarded as a parametric model with a number of parameters approaching to the number of observations, the effect of excessive degrees of freedom may be manifested in the figures of Table 3.

Figures 5 and 6 illustrate how the goodness-of-fit measures in both the calibration and the holdout sample vary as the smoothing constants h and μ for the continuous and discrete variable, respectively. As expected, the lesser the smoothing is ($h \rightarrow 0$ and $\mu \rightarrow 1$), the better the fit is for the calibration but the worse for the holdout sample. Also, in the calibration sample, the fit degrades as the smoothing is increased.

3.5 Market Responses

Table 4 lists market responses, percent change in share by feature and display activities. They are computed by changing the values of the chosen attribute for the corresponding brand keeping the other variables unchanged and aggregating the predicted probabilities. The discrepancy between NDE and MNL needs further investigation, especially Ssips figures of NDE where display rarely occurred according to Table 1. Some discussion on these response figures by MNL will be found in the next section.

Table 4: Percent Share Change due to Own-Brand Promotions by NDE and MNL

	FEATURE		DISPLAY	
brand	NDE	MNL	NDE	MNL
Hi-C	23.1%	28.7%	23.9%	41.5%
KoolAid	59.9%	35.6%	42.2%	51.9%
Ssips	5.8%	16.1%	1.3%	23.3%

3.6 Marginal Probability

Further insight is obtained by focusing onto one or two variable dimensions of the estimated density function. The first example is a plot of the share for Hi-C v.s. loyalty of each brand by aggregating all other dimensional variables, which is shown in Figure 9. Although, Hi-C

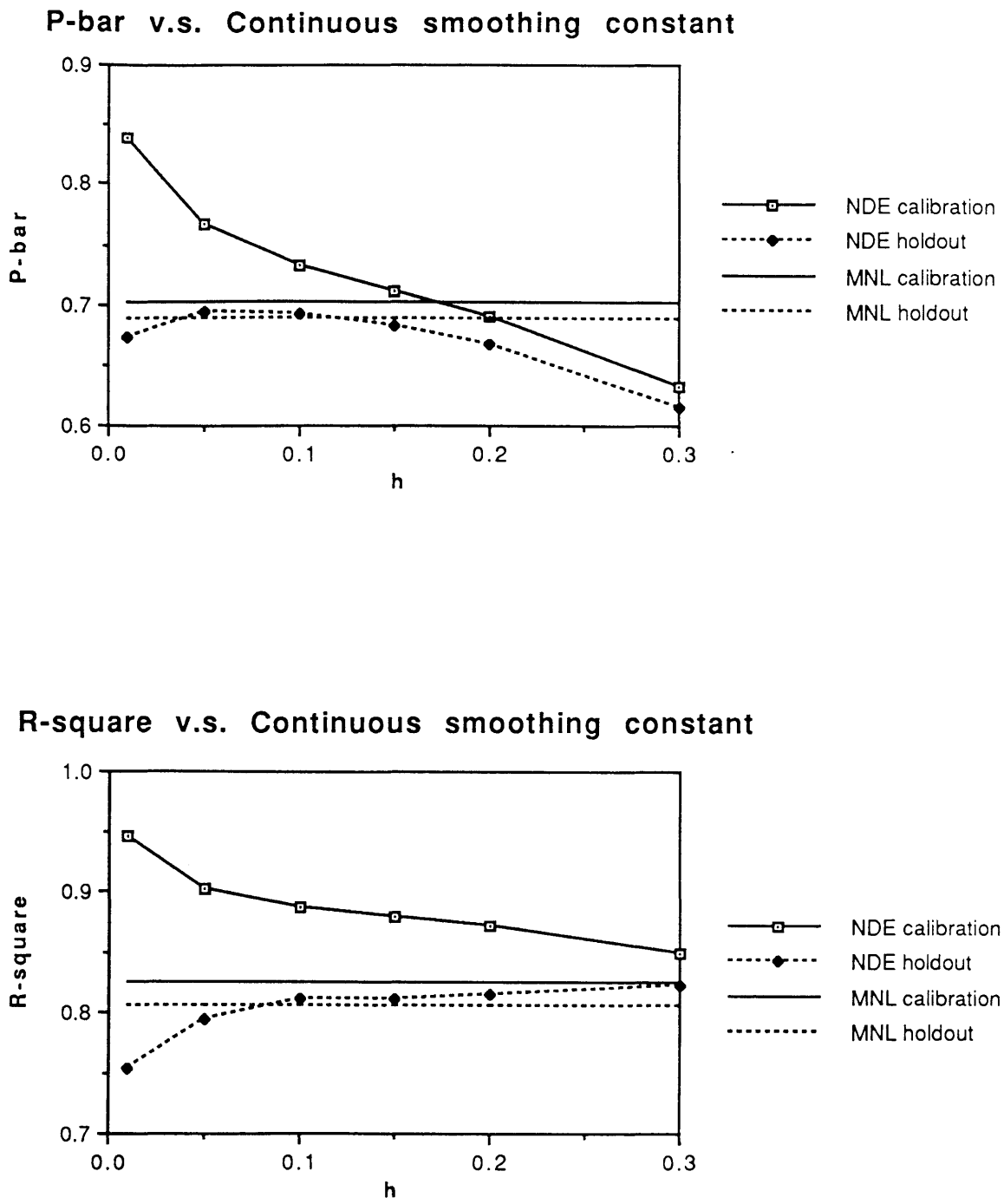


Figure 5: Goodness-of-Fit v.s. the Continuous Smoothing Constant h

As the smoothing is decreased (h is decreased), goodness-of-fit improves in the calibration. In the holdout, R^2 degrades while $\overline{P}$ improves.

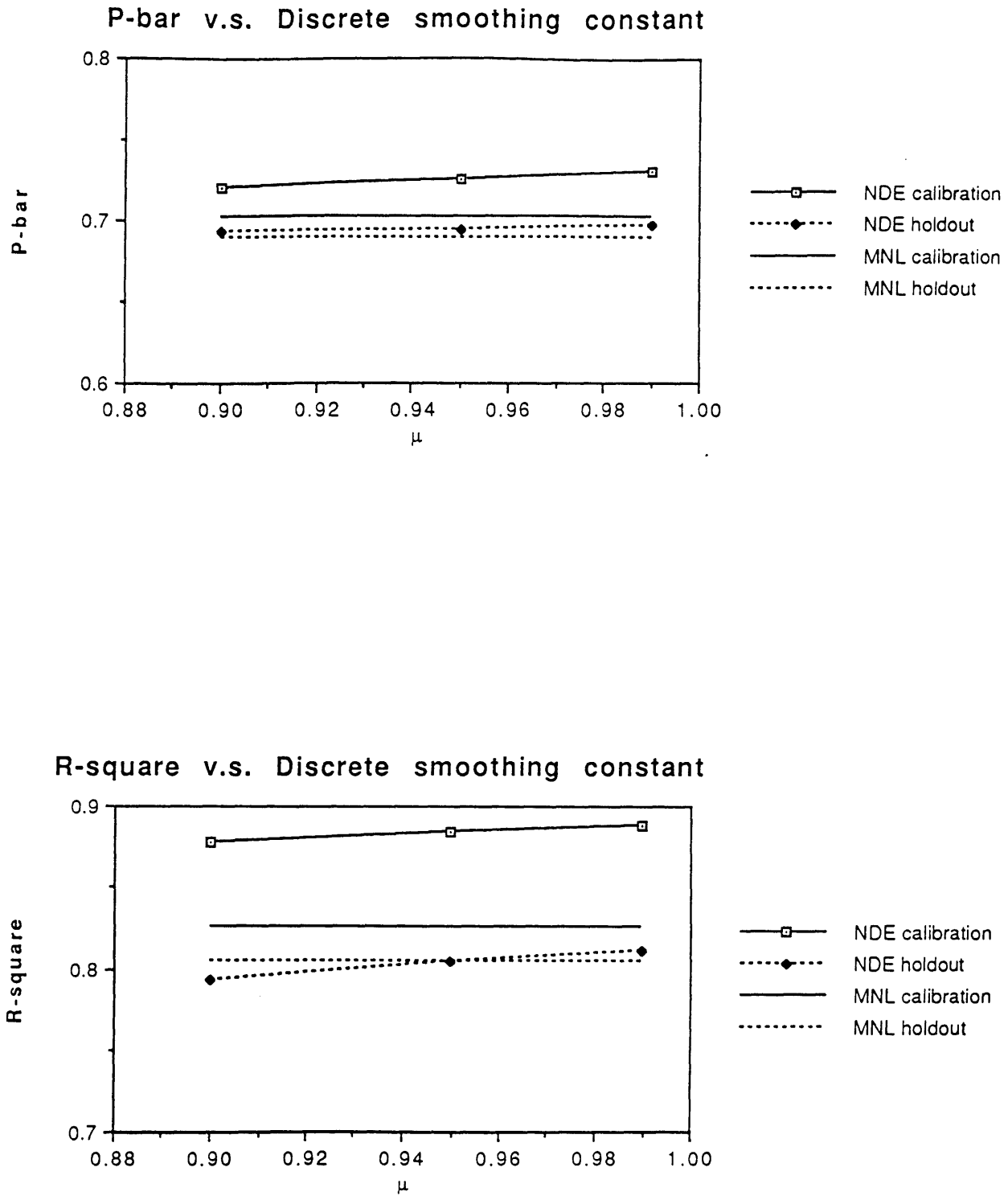


Figure 6: Goodness-of-Fit v.s. the Discrete Smoothing Constant μ

As the smoothing is decreased (μ is increased), goodness-of-fit improves both in the calibration and holdout.

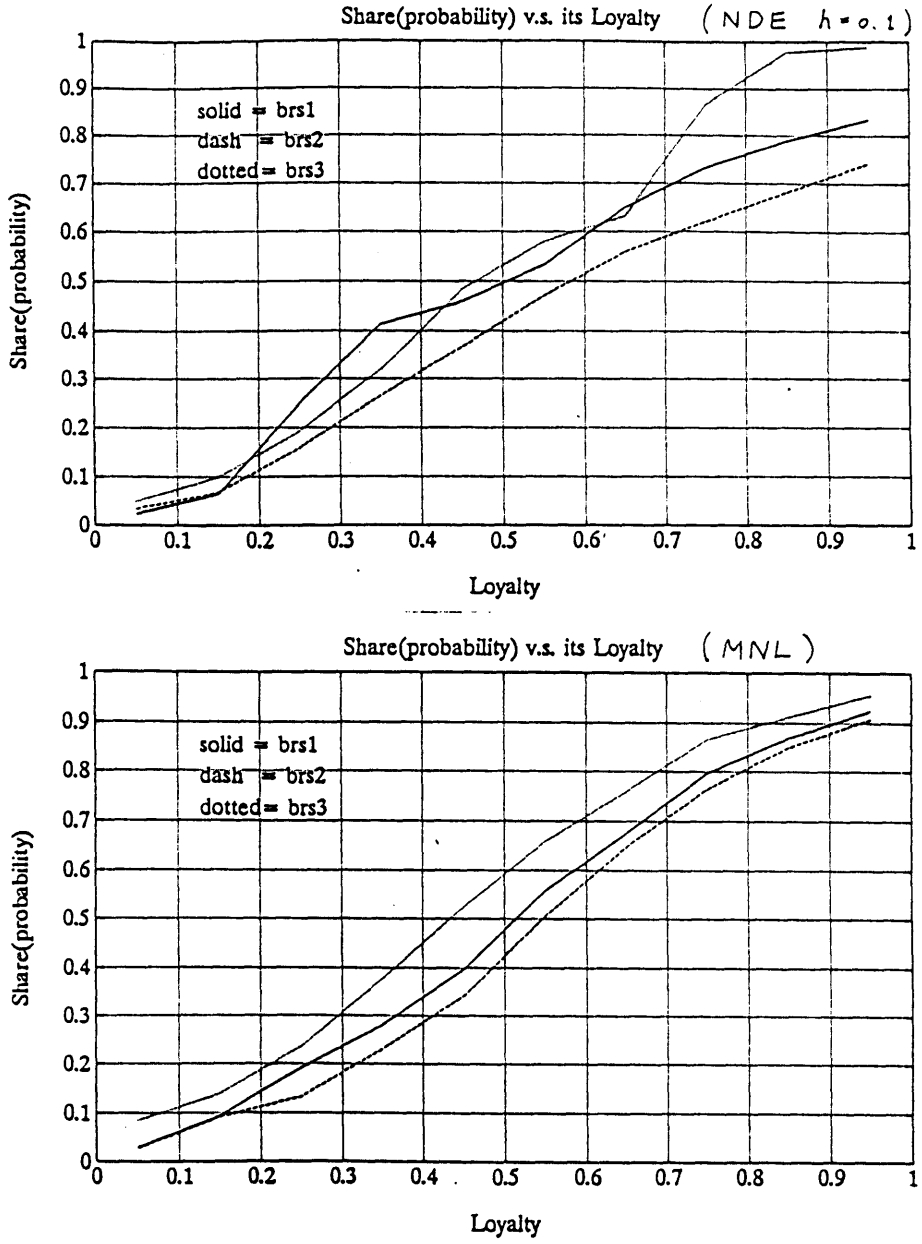
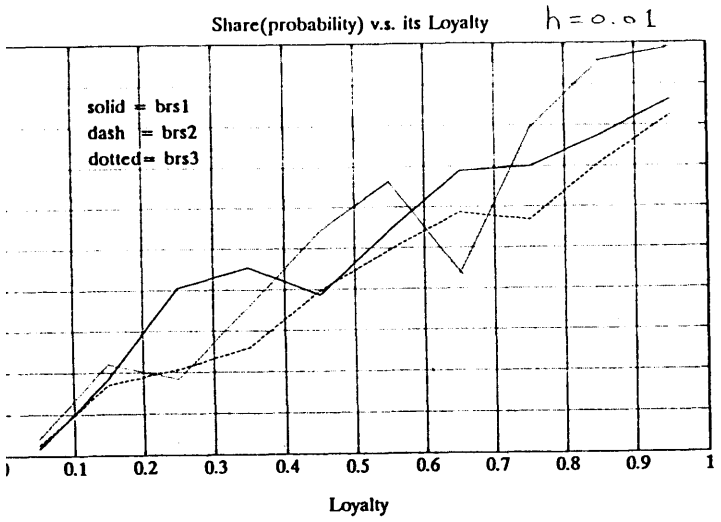
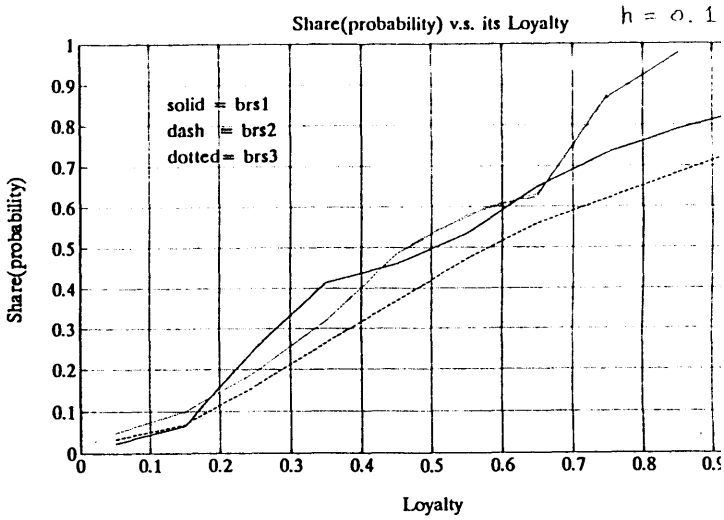
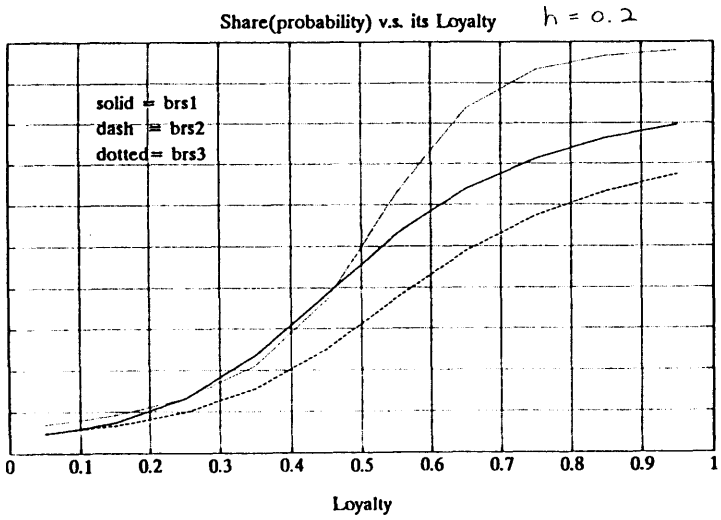


Figure 7: Marginal Probability v.s its Own Loyalty by NDE (h = 0.1) and MNL

Casual Comparison of the marginal probability curves suggests that they look rather similar. But NDE version may offer further improvement over MNL version.







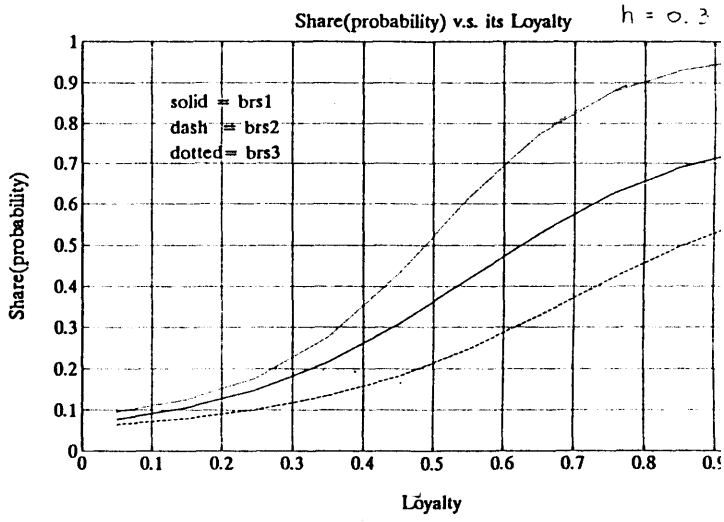


Figure 8: Marginal Probability v.s Loyalty with Different Smoothing Constants

The curve becomes smoother as the smoothing constant, h, is increased. h=0.3 seems to be over-smoothed while h=0.01 looks under-smoothed.

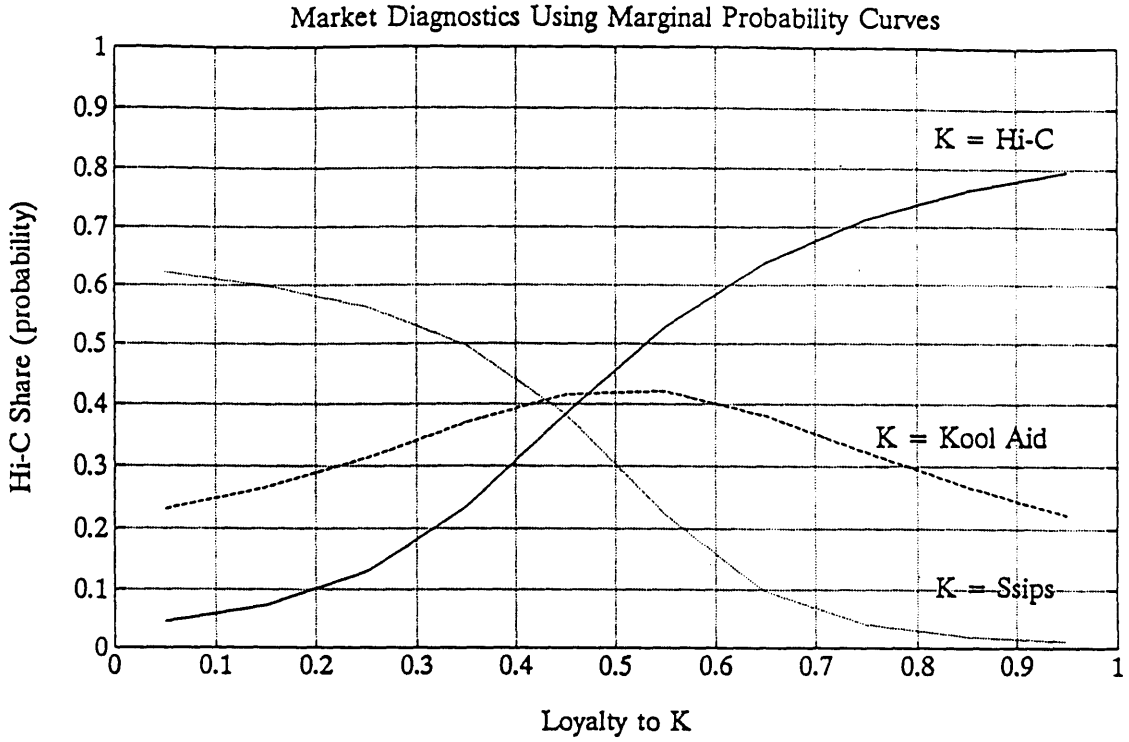


Figure 9: Marginal Probability of Hi-C v.s Loyalty of K by NDE

As expected, increasing Hi-C loyalty and/or decreasing Ssips loyalty is associated with increasing Hi-C share. But surprisingly, increasing KoolAid loyalty (for low KoolAid loyalty) is associated with increasing Hi-C share.

share increases in loyalty of Hi-C and decreases in loyalty of Ssips as expected, the non-monotonic curve corresponding to loyalty of KoolAid is rather surprising. It implies that increasing KoolAid loyalty (for low KoolAid loyalty) is associated with increasing Hi-C share, which is counter-intuitive. One possible explanation could be that, both Hi-C and KoolAid being famous national brands as opposed to the relatively unknown price brand Ssips, there exists a loyalty process across the two former brands which constitute a high quality segment.

The second example is shown in Table 5, a 2×2 table of KoolAid share for various combinations of feature and display of KoolAid by again aggregating other dimensions. This indicates a strong interaction effect of the two promotion variables because the share would be only 0.265 had there been no interaction.

Table 5: Interaction of Feature and Display on KoolAid Share

	no display	display
no feature	0.149	0.176
feature	0.225	0.414

4. DISCUSSION

4.1 Comments on MNL

The independence from irrelevant alternatives (IIA) property potentially influences all market responses such as elasticities of MNL. For example, disaggregate cross elasticities of the probability of alternative i with respect to the k-th attribute of alternative j≠i is independent of i because its expression corresponding to the n-th purchase incident is

$$E^{P_n(i)}_{x_{njk}} = \frac{\partial P_n(i) / P_n(i)}{\partial x_{njk} / x_{njk}} = \left[\delta_{ij} - P_n(j)\right] x_{njk} \beta_k$$

where β_k = coefficient of the k-th attribute.

Thus, MNL imposes uniform cross elasticities at the disaggregate level. Now, define aggregate share elasticity (Ben-Akiva and Lerman 1985) as

$$E_{x_{jk}}^{\bar{P}(i)} = \frac{\partial \bar{P}(i) / \bar{P}(i)}{\partial x_{jk} / x_{jk}} \qquad \text{where} \quad \bar{P}(i) = \frac{\sum_{n=1}^{N} P_n(i)}{N}, \quad \text{and} \quad \frac{\partial x_{njk}}{x_{njk}} = \frac{\partial x_{n'jk}}{x_{n'jk}} = \frac{\partial x_{jk}}{x_{jk}}.$$

Although the uniformity property does not directly carry over to the aggregate elasticities, there is still a heavy influence from the reciprocal of the average shares since the aggregate elasticity of alternative i to the k-th explanatory variable of alternative j, x_{jk} can be written as

$$E_{x_{jk}}^{\bar{P}(i)} = \frac{\beta_k}{N \bar{P}(i)} \sum_{n=1}^{N} P_n(i) \left[\delta_{ij} - P_n(i) \right] x_{njk}.$$

The effect can be readily observed in the market response figures of MNL in Table 4. The percent share change due to feature and display are approximately inversely proportional to their shares, exhibiting the denominator effect.

Also for attributes associated with common coefficients across the alternatives, only their differences between the alternatives are relevant on a choice probability, irrespective of their absolute levels. The implication is two fold. Let d be the difference of a particular attribute, say price, between brs 1 and brs 2 as $d = P_2 - P_1$. Then,

(1) Price increase of δ by brs 2 produces the same effect as price reduction of δ by brs 1 because

$$d + \delta = (P_2 + \delta) - P_1 = P_2 - (P_1 - \delta)$$

This is implicitly related to the property of the uniform cross elasticity.

(2) If prices of both brands are raised by the same amount δ, then its net effect is zero and their choice probabilities are unchanged if the other attributes stay the same since

$$d = P_2 - P_1 = (P_2 + \delta) - (P_1 + \delta)$$

The lines in the marginal probability v.s. its own loyalty plot for the three brands by MNL shown in Figure 7 are near parallel with different absolute levels unlike that by NDE. This is due to the common loyalty coefficient across the alternatives in the logit formulation, which exemplifies the above phenomena.

4.2 Comments on the Nonparametric Density Estimation

The main advantage of NDE is that very few assumptions are required unlike MNL. And its concept is very straightforward since the joint PDF is a cumulation of the past history and can be interpreted as a "smooth histogram". Derived from this are the marginal probabilities on which various exploratory analyses can be conducted. Therefore, NDE facilitates analysts to examine data itself, while it is often the case in MNL to see raw data only in the form of a time series of shares.

Its structureless assumption makes the method extremely flexible to conform to any shape of distribution. Since the resulting density function is based solely on data fed in, its prediction for the holdout sample is excellent in the range of explanatory variables where observations are repeatedly made in the calibration. The flip side is that, due to the lack of a priori structure, its performance in prediction is relatively poor, if not impossible, for what has not been encountered before. In other words, an outcome of extrapolation must be interpreted with caution.

Similarly, the market response for Ssips due to its display activity needs further investigation since its occurrence happens mere 5 times or only 0.2% of the data. This does not imply that the figure by NDE is incorrect, but care must be exercised. In contrast, the same figure for Ssips derived from MNL (Table 4) is mainly driven by the response of Hi-C to its own display activity because the display coefficient, which is common across alternatives, is estimated largely from data of Hi-C which displays 16 times as often as Ssips. Thus, unless magnitude of the response to its own display activity is similar between Hi-C and Ssips, MNL estimates cannot be relied upon either. If alternative specific display coefficients are utilized to overcome this difficulty, it is likely to encounter the small sample instability in estimating the Ssips' coefficient. This will bring MNL to the same situation as NDE.

An even more interesting case is the prediction of shares when a new product is introduced. The NDE is not able to answer this because of the absence of the previous history. The MNL will generate a purchase probability of the new brand if its values of the attribute variables are

provided. The prediction will be based on IIA, that is, the relative utility of the new product determines the level of the share it acquires, and the remaining brands are pushed back in proportion to their pre-entry levels. This may not be exactly true. Therefore, it is either "no answer" by NDE or "possibly inaccurate answer" by MNL.

4.2.1 Computational requirement with respect to time and space

Two main disadvantages of NDE over MNL are (1) large computational requirement in terms of time and storage, and (2) large sample requirement. In this study, the run time for NDE method took about 2,000 seconds against 30 seconds by MNL using a 16MHz 386 machine. This is a factor close to 70. Partially compensating for the speed problem is NDE's ability to incrementally update the density function. Because the density function is a sum of the appropriately normalized kernels, it can be revised with ease as new data becomes available. The MNL does not possess this property since parameters would usually be re-estimated from scratch every time new data arrives. A need for large storage space to save the joint PDFs in the form of discrete approximation is another matter, however. Here, there exists three such PDFs, each with 57,600 cells, even after additional assumptions were imposed in an effort to reduce this size. These assumptions are (1) omitting PRICE attribute variables, and (2) introducing the loyalty variable which is a parsimonious aggregation of more fundamental observation variables.

4.2.2 Large data requirement in higher dimensions

While these computational limitations may not be crucial to the utilization of the method because of advances in computer technology, the real bottleneck is in the necessity of a huge amount of data for reliable estimations, especially in a PDF of larger dimensions. Because the number of attribute variables increases linearly with the number of alternatives at hand, the number of cells becomes enormous and even the largest databases result in few observations per cell. Then, the estimation becomes unstable and inefficient. The following simple examples excerpted from Silverman (1986) demonstrate why the common sense in low dimensions completely breaks down in higher dimensions. This exponential growth of required sample size as the number of dimension is often referred to as "curse of dimensionality" and is a well known phenomenon.

Example 1: Approximating the distribution of a tail whose magnitude is less than 1% of the maximum value is fairly irrelevant in estimating an accurate density for one dimensional case. But, consider estimating a density function whose underlying true distribution is multivariate

independent normal with mean 0 and variance 1. Then the probability of points falling in the region x such that $f(x) \leq 0.01 \times \sup f$ is less than 0.5% in 1 dimension and about 1% in 2 dimensions. But in 10 dimensions, about 50% of all points reside in this region.[2] Let us take a look at this surprising fact from a different angle. While 90% of sample lies between ±1.6 in 1 dimension, only 1% lies in the region whose distance is less than 1.6 from the origin in 10 dimensions.[3] Therefore, estimation of the tail is crucial.

Example 2: Consider an even simpler case where the underlying density function is a uniform hypercube of edge lengths equal 1 centered at the origin. In one dimension, clearly 1% of the sample occupy the length of 0.01 on average. Yet, the same 1% fill in the hypercube of edge lengths equal to 0.63 in 10 dimensions because $(0.63)^{10} \approx 0.01$.

Both examples can be described as an "empty space phenomenon", in which most points are observed where the magnitude of f(x) is very small. This makes for the need of the enormous number of observations in higher dimensions.

5. CONCLUDING REMARKS

In the previous section, two contrasting methods are compared in a brand choice context. One side is represented by the highly parametrized MNL, and the opposite end by NDE. Advantages and disadvantages of each method are discussed in detail. If we summarize the two in terms of modeling criteria as in Table 7, the difference becomes even clearer.

[1] **Predictability**: How well does a model perform in prediction by extrapolation as well as interpolation?

---

[2] This can be derived as follow.

$$\frac{f(x)}{f(0)} = \exp\left[-\frac{1}{2}x^t x\right] \sim \exp\left[-\frac{1}{2}\chi^2_d\right]$$

$$\frac{f(x)}{f(0)} = 0.01 \Rightarrow \chi^2_d(\text{prob}) = 9.21$$

[3] This is since $\chi^2_{10}(0.99) = 1.6^2 = 2.56$

[2] **Robustness to underlying assumptions**: How well does it perform if underlying assumptions are violated?

[3] **Descriptive capability**: How well does it provide an underlying structure for understanding the phenomenon?

[4] **Adaptability**: How easy is it to update the model?

[5] **Operating Characteristics**: Sample size requirement, computational time

Table 7: Evaluation of the two models by the five criteria

Criterion	NDE	MNL
Predictability	-	+
Robustness to underlying assumptions	+	-
Descriptive capability	-	+
Adaptability	+	-
Operating characteristics	-	+

Any model can be positioned between these two extrema on the continuum line of varying degree in parametrization, which is shown in Figure 10. The MNL in our study had only five parameters, and incorporated stochastic utility maximization with a doubly exponentially distributed disturbance term. It is surprising that such a parsimonious model can compete reasonably well against NDE which could be regarded as having many more number of parameters. The probit model is located slightly away from logit toward less parametrization, and hence exhibits more flexible structure. But it is necessary to estimate more parameters including variance-covariance matrix of the disturbance term. Although NDE demonstrated here may not be the pure ideal nonparametric method due to assumptions with respect to selection of the attribute variables, shape of the kernel, and i.i.d. observations, it is exceedingly nonparametric in the traditional sense. In reality, one would like to opt for a model satisfying all five criteria reasonably well, which ought to be somewhere between these two extrema.

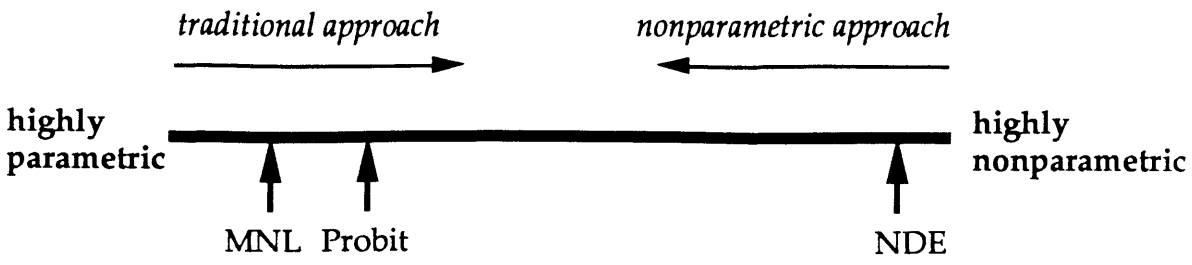


Figure 10: Models along the continuum line of varying degree in parametrization

For future research, there exists two major directions in extending the nonparametric methods. One is to develop diagnostic techniques for parametric models using nonparametric methods. In other words, is it possible to test some parametric model assumptions by relaxing them with a nonparametric method? One such attempt is a study done by Abe (1991), where an empirical distribution of the random component of MNL is constructed using a nonparametric regression and then compared against the theoretical one (logistic function) to check the distributional assumption of logit. The other is to investigate a class of so called semiparametric methods, which are a compromise between parametric and nonparametric methods, so that they are located somewhere in the middle of the line in Figure 10. These two areas should complement each other during the course of building good models for achieving one's objectives.

Although the theoretical aspects of nonparametric methods have been extensively covered in the past decade, application studies hardly exist in the literature. For other empirical studies of a nonparametric density estimation method in the marketing field, see Bumbaca (1988), Rust (1988), and Donthu & Rust (1989). UPC scanners generate an enormous amount of data across the nation everyday, and the benefits from its efficient analysis and full utilization are large. Though some might see this data glut as one big computational problem, it must be regarded as a golden opportunity. While in most experimental disciplines, availability of data is the bottleneck, here is a rare situation where there is an abundance of data. In this respect, marketing modelers should consider fresh, if not radical, approaches without being constrained in the traditional view of data analysis.

APPENDIX

Asymptotic Properties of Kernel Density Estimator

First, sufficient conditions for the pointwise strong consistency of the estimated function using n points, f_n, are, (i.e. $f_n \to f$ as $n \to \infty$ almost surely)

[1] Let K be the class of all Borel-measurable bounded real valued functions, K(t), of d-dimensional vector **t**, such that

(a) $\int K(t)\, dt = 1$, (b) $\int |K(t)|\, dt < \infty$, (c) $\|t\|^d |K(t)| \to 0$ as $\|t\| \to \infty$,

(d) $\text{Sup}\,|K(t)| < \infty$, where $\|t\|$ is the Euclidean norm of **t**.

[2] $h_n \to 0$ as $n \to \infty$

[3] $n(h_n)^d \to \infty$ as $n \to \infty$

[4] f(t) is continuous for each t.

Second, pointwise asymptotic normality of f_n was shown by Parzen (1962) as

$$\frac{f_n - E f_n}{\sqrt{V(f_n)}} \to N(0,1) \quad \text{as } n \to \infty$$

where $V(f_n)$ is the asymptotic variance of f_n expressed as

$$V(f_n) = \frac{1}{n h^d} f \int K^2(t)\, dt$$

Third, MISE decreases as $n^{-4/(d+4)}$ at best for the case of the previous section. This can be seen by substituting the optimal h of (4) into equation (3).

Estimating an Additive Nonparametric Utility Function in Logit Models of Brand Choice by Utility Residual Method

Makoto Abe

Operations Research Center
M.I.T.
Cambridge, MA 02139 USA

M. I. T. Doctoral Dissertation, Part II

June 1991

OVERVIEW

Marketing managers in packaged goods companies now have access to great quantities of scanner data. This suggests the possibility of nonparametric methods. These can take special advantage of large datasets to provide new modeling flexibility. A previous nonparametric approach to a brand choice model described in Part I, however, has shown that the direct approach of nonparametric density estimation requires even larger databases than may be reasonably expected from scanners.

The present paper takes a mid path that tries to overcome such difficulties while maintaining flexibility in model specification by pursuing a semiparametric approach. The proposed utility residual method (URM) for the multinomial logit keeps the doubly exponential logit distributional assumption while expanding the structural freedom of the systematic utility component. This is done by using a one dimensional nonparametric function for each explanatory variable in additive utility form. The key idea is to develop a residual for the utilities in the multinomial logit. The utility residual contrasts to the ordinary residual of a discrete choice model, which is the difference between observed binary choices and predicted probabilities. The utility residuals provide much greater insight into market response functions and behavioral phenomena.

The validity and usefulness of URM is verified in a simulation study and in applications to two scanner databases. In the simulation, a utility structure is recovered nonparametrically from a pre-specified additive nonlinear utility function in a logit model. The two scanner data applications illustrate the development of nonlinear additive utility functions and their marketing implications.

1. INTRODUCTION

Recent advances in scanner technology have made available large databases of individual purchase records and opened up a whole new direction in marketing. The issues of interest cover a wide range from brand choice, purchase quantities, interpurchase timing to behavioral theories of price, advertising, and promotion response as well as repeat purchasing. In studying these, many models have been created to address specific questions to databases which contain enormous amount of information. Most of these models are parametric in nature, in other words, a particular functional form with some number of unknown parameters is assumed based on theories and/or data. The parameters are then estimated from the data, various tests are made, and conclusions are drawn. For example, a multinomial logit (MNL) brand choice model is founded on stochastic utility maximization with a specific assumption on the random component of the utility. From this, a choice probability can be derived to have a particular functional form of the deterministic utility, which is usually expressed as linear-in-parameters of relevant explanatory variables.

In many parsimonious MNL models applied to scanner panel data, one frequently observes t-values for some coefficients such as loyalty to be as high as 40, because large databases with thousands or even tens of thousands of observations are quite common. (Guadagni and Little 1983) Yet, obtaining such high t-values is meaningless if the underlying model is incorrectly specified, in which case the parameter estimates will be biased and subject to misinterpretation. A traditional approach to the model specification is to propose an initial model based on existing theories, common sense, and experiences and prior knowledge of the analyst, and then keep refining it with various specification tests. Such a trial and error process is not only time consuming, but also results in the final model which varies greatly depending on the subjective judgements of the analyst. With abundant data collected by scanners and appropriate statistical techniques, perhaps we can afford to let the data do more of the work of specifying the model structure than has been possible heretofore.

For a brand choice model, this idea is pushed to an extreme of empiricism by the nonparametric density estimation in Part I of this dissertation, where a conditional probability of brand choice y given a set of marketing mix variables is calculated from a multidimensional joint probability density function, f(x,y), as

$$\text{(1)} \qquad P(y \mid x) = \frac{f(x,y)}{f(x)} = \frac{f(x \mid y)\, f(y)}{\int f(x \mid y)\, f(y)\, dy} .$$

The density, f(x,y), is estimated nonparametrically by a kernel method (Silverman 1986). Conceptually, it can be interpreted as a huge multidimensional smoothed histogram or cross tab. Because the nonparametric density estimation (NDE) method simply compiles a history of past observations to predict the future, it does not involve a parametric specification of the utility nor a distributional assumption unlike MNL. In fact, the whole notion of the utility maximization does not exist at all.

Two notable disadvantages of NDE are, [1] a large computational requirement, and [2] a large sample requirement. Part I finds a computational factor of close to 70 for NDE over MNL for a dataset of about 1000 observations, 3 alternatives with two continuous and two binary explanatory variables. However, as more powerful desktop computers become available, the computation may not be the limiting factor. The real bottleneck is the second one, i.e. the necessity of an enormous amount of data to reliably estimate the density function, especially for larger dimensions. Because in NDE the number of attribute variables increases linearly with the number of alternatives at hand and the number of cells increases exponentially with the number of attribute variables, even the largest databases result in only a few observations per cell. Eventually, the estimation becomes unreliable. The exponential growth of required sample size with the number of dimensions is often referred to as the "curse of dimensionality", and is a well-known phenomenon. (Friedman and Stuetzel 1981, Silverman 1986) Other applications of nonparametric methods in marketing can be seen in Rust (1988) and Rust and Donthu (1989).

Such a difficulty encountered in the NDE method leads us to consider a step back from such extreme empiricism by taking a mid approach that employs semiparametric methods. In other words, a basic foundation of a model is built around well-established theories, while data specifies the remaining structure.

In the present paper focusing on a brand choice model, the model is founded on the MNL framework, i.e. stochastic utility maximization with a doubly exponentially distributed random utility component, but the systematic utility is specified nonparametrically. In particular, the utility is specified as a sum of one dimensional nonparametric functions of relevant explanatory variables. Hence, the choice probability of alternative j is expressed as

(2) $$P_j = \frac{e^{v_j}}{\sum_k e^{v_k}} \qquad \text{where} \qquad v_j = \sum_p \phi_p(x_{jp})$$

where x_{jp} is the p-th attribute variable for alternative j and $\phi_p(\cdot)$ is a one dimensional nonparametric function for the p-th attribute.

Semiparametric methods in the literature can roughly be classified into two groups, depending on whether a distributional assumption or a model specification is nonparametrized. Relaxing random term distributional assumptions such as normality, symmetry, and homoskedasticity, is rather beneficial in practice for obtaining consistent parameter estimates. In econometrics, these techniques are called distribution-free methods and there exist various papers. (Manski 1975, 1986, 1989, Cosslett 1983, Duncan 1986, Stoker 1986, Han 1987, Klein and Spady 1988) The latter group, which imposes certain parametric distributional assumptions while relaxing a model specification is relatively new, and has been studied mainly in the field of statistics (Breiman and Friedman 1985, Hastie and Tibshirani 1986 1987 1990, Matzkin 1989).

There are several reasons for building our semiparametric model around MNL. One is that MNL has been confirmed to perform quite well as a brand choice model, especially in cross validations, with various data sets, and has become a standard benchmark for comparing choice models in the past decade. As with any model, MNL involves certain assumptions which may or may not reflect reality. But, its proven record in field studies supports the robustness of the MNL assumption in practice. Two properties of MNL might explain its operational successes. First, when the conditional choice probability of j given a vector of covariates x, $P(j|x)$, is expressed by the likelihood function and a prior using Bayes rule, it resembles the general MNL formulation. Furthermore, if the priors are the same and covariates are normally distributed with identical covariances but different means across alternatives, the expression becomes Fisher's discrimination method and reduces to MNL with linear-in-covariates utility form. Second, among a class of multinomial distributions whose probabilities are functions of a linear predictor of covariates, $\eta = x'\beta$, a logit function, $\eta = \log[p/p(1-p)]$, is referred to as a canonical link by the terminology of the Generalized Linear Models (GLM) and possesses a sufficient statistic for β. (McCullagh and Nelder 1989) Since the derivation is not documented in their book, it is shown in Appendix D.

The second reason for choosing MNL is its underlying behavioral mechanism of stochastic utility maximization, where a consumer chooses an alternative which has the highest utility but the utility has a random component. Possible sources for such stochastic disturbance are

unobserved attributes, unobserved heterogeneity, measurement error, and imperfect (instrument and/or latent) attributes. Therefore, the stochastic component accounts for uncertainty from the consumers' side as well as model deficiency from the analysts' side.

Third, because there exists an explicit analytical expression for a choice probability, MNL is mathematically tractable and easy to build variants and extensions.

Finally, various diagnostic tests are available to check for the distributional assumption of MNL including IIA tests. (McFadden, Tye, and Train 1976, Hausman and McFadden 1984, McFadden 1985) In Abe's paper (1991) on a general kernel regression called moving ellipsoid method, he illustrates an application which permits a visual comparison between the empirical and theoretical (i.e. logistic) distribution of the stochastic component. Thus, by using these techniques, we can make sure that the current method is valid by examining whether the MNL distributional assumption is violated or not. In addition, because our method constructs an empirical random distribution at each iteration, a consistency check for the logit assumption can be performed by monitoring its deviation from the theoretical distribution.

A nonparametric utility specification of the form (2) can potentially unveil various interesting marketing phenomena such as nonlinearities and asymmetries in explanatory variables. Uncovering such structures using the conventional linear-in-parameters approach requires skilled analysts and elaborate modeling schemes in which special variables and sequences of models are created to facilitate hypothesis testing. The method to be described does not involve such an extensive iterative process to reach an acceptable model specification. Furthermore, its output, nonparametric plots with partial utility residuals, can help identify outliers and influential points and visually communicate with non-technical managers who may be able to provide better marketing interpretations than data analysts. Hence, relaxation of model specification is of great value for investigating managerial implications as well as substantive issues in marketing.

In terms of the organization of this paper, Section 2 explains our method, which we shall refer to as the utility residual method (URM). It consists of two iterative steps. One infers residuals in utility from discrete choice data, and the other constructs the additive nonparametric utility function from the residuals. Section 3 is a simulation study to illustrate how well URM recovers a pre-specified nonlinear utility structure. In Section 4, applications of URM to two scanner databases are demonstrated, and their marketing implications are discussed. In the second database, URM is evaluated both in calibration and holdout sample

against linear-in-parameters MNL and NDE. This illuminates the relationship between predictive ability and strength of assumptions or equivalently the degrees of freedom. Section 5 summarizes the paper and discusses possible extensions .

2. A SEMIPARAMETRIC METHOD USING UTILITY RESIDUALS

2.1 Overview

The choice probability of alternative j by the standard linear MNL can be expressed as

$$P_j = \frac{e^{V_j}}{\sum_k e^{V_k}} \qquad \text{where} \qquad V_j = \sum_p \beta_p x_{jp}$$

where x_{jp} denotes the p-th attribute value for alternative j. The objective of the utility residual method (URM) is to obtain a MNL discrete choice model with flexible utility structure as follows.

(2) $$P_j = \frac{e^{V_j}}{\sum_k e^{V_k}} \qquad \text{where} \qquad V_j = \sum_p \phi_p(x_{jp})$$

where $\phi_p(\cdot)$ is a one dimensional nonparametric function.

We consider additive separability in marketing variables for three reasons. First, it is a natural generalization of the linear model. Second, an interaction term can readily be incorporated by creating a new variable which multiplies relevant variables together and extending the p-index. Third, if we start replacing several 1-dimensional functions $\phi_p(\cdot)$ by a more general multidimensional nonparametric function $\phi_{pq..}(\cdot,\cdot,..,\cdot)$, we are likely to encounter the "curse of dimensionality" problem (Stone 1985, Silverman 1986). That is, an exponentially increasing sample size will be required as the number of dimensions grow to maintain reliable estimation as shown in Part I.

In ordinary least squares, the correctness of the linear model is often diagnosed by examining the residuals, $e_n = y_n - \Sigma \beta_p x_{np}$, on a partial residual plot. (Mosteller & Tukey 1977) In fact, the Alternative Conditional Expectation (ACE) method by Breiman and Friedman (1985) and the Generalized Additive Models (GAM) by Hastie and Tibshirani (1986, 1987) obtain nonparametric regression of the following kind from the residuals, which is analogous to the utility in equation (2). (see Figure 1)

$$y = \sum_p \phi_p(x_p)$$

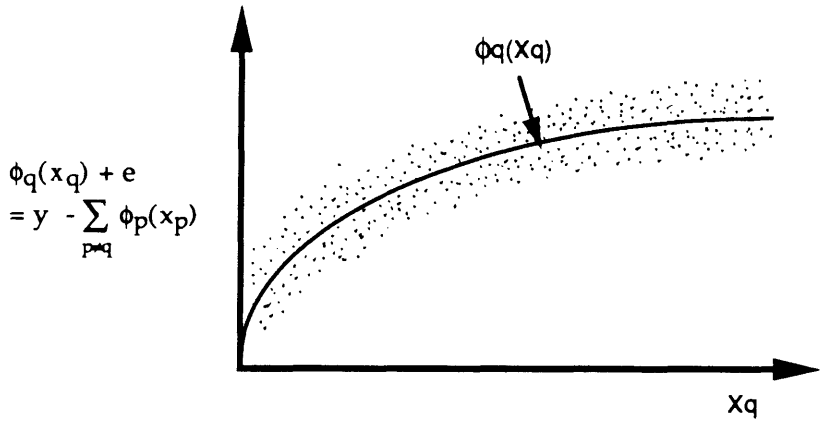


Figure 1: OLS residuals against explanatory variable q

A vertical difference between a point and the curve, $\phi q(xq)$, is a residual.

The difficulty of these approaches for discrete choice models is that v_j is not directly observable. That is, we do not have a straightforward way to find e_{jn} in (3).

(3) $$e_{jn} = v_{jn} - \sum_p \beta_p x_{jnp}$$

In the following subsections, a concept of utility residuals is introduced, and a computational method, Utility Residual Method (URM), for finding ϕ_p is presented.

2.2 Utility Residual Method

The method assumes [a] the additive separable utility function, and [b] that the MNL distributional assumption, i.i.d. doubly exponential distribution, of the disturbance term holds for discrete choice data in question. In MNL, the choice probability of alternative j can be expressed as follow.

(4) $$P_j = \frac{e^{v_j}}{\sum_k e^{v_k}} = \frac{1}{1 + e^{-w_j}}$$

(4a) where $$w_j = v_j - \ln \sum_{k \neq j} e^{v_k}$$

Hence P_j is a logistic function of w_j as the name logit suggests. Equation (4) can also be derived from stochastic utility maximization with the doubly exponential random term. This derivation in Appendix A shows that $P_j = \text{prob}(\xi < w_j)$, where ξ is logistically distributed.

If an estimate of w_j is known (e.g. from the linear model), then $\hat{P}_j$, which is an empirical value of P_j, can be obtained from nonparametric regression of y_j on w_j as $\hat{P}_j(w_j) = E(y_j \mid w_j)$, where y_j is 1 if alternative j is chosen and 0 otherwise. See Figure 2.

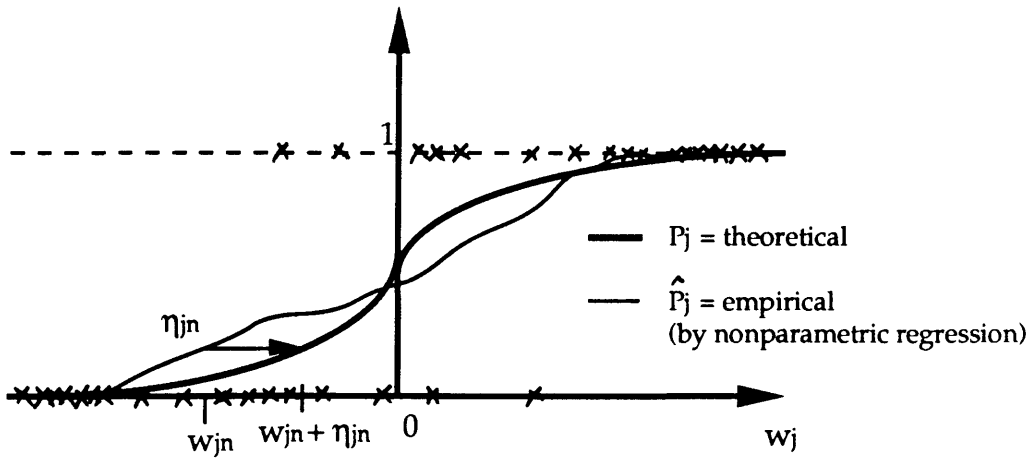


Figure 2: Theoretical logistic function and the empirical probability plot, $\hat{P}_j$

A residual in utility is a horizontal difference between the two curves.

Since the regression involves just a single explanatory variable, several alternative methods exist as the nonparametric regression. In our case, a kernel estimation method is used because it possesses an attractive mathematical property of a consistent estimator and can be readily extended to higher dimensions. Details of the nonparametric regression appear in Section 3.3.

If, [a] the model is correctly specified (i.e. v_j is correct), and [b] the disturbance is doubly exponential, then $\widehat{P}_j(w_j)$ should be close to $P_j(w_j)$ and approach it asymptotically as the amount of data increases. Maintaining assumption [b] of the doubly exponential disturbance, we take the discrepancy between $\widehat{P}_j$ and P_j as a measure of the misspecification in v_j, i.e. the violation of [a].[4] Checking the difference between the empirical and theoretical probability plot is somewhat analogous to the normal probability plot in linear regression. In the latter case, the normal probability plot is used simply to check the normal distribution assumption of the error term and cannot be used to infer model adequacy of $X\beta$. However, we can do more in the logistic case. Due to the discrete nature of the response with expectation p and variance p(1-p), the error distribution is directly affected by the model specification. (Landwehr, Pregibon, and Shoemaker 1984) This is illustrated in Section 3.2 by simulation study.

For the n-th observation, we calculate w_{jn} from equation (4a) and define a utility residual as the amount, η_{jn}, that must be added to w_{jn} to convert P_j into $\widehat{P}_j$, i.e.

(5) $$\widehat{P}_j(w_{jn}) = P_j(w_{jn} + \eta_{jn}) = \frac{1}{1 + e^{-(w_{jn} + \eta_{jn})}}$$

In Figure 2, η_{jn} is the horizontal distance between $\widehat{P}_j$ and P_j at $\widehat{P}_j(w_j)$. Using (5), it is expressed as,

$$\eta_{jn} = -w_{jn} - \ln\left\{\frac{1}{\widehat{P}_j(w_{jn})} - 1\right\}$$

[4] As a technical note, a smoothing constant for the nonparametric regression (whatever is utilized) must chosen such that the empirical distribution $\widehat{P}_j$ becomes monotonically increasing in w_j to avoid a double value in the discrepancy.

Starting from the linear model, we shall iteratively estimate $\phi_q(\cdot)$ in such a way as to reduce the residuals, η_j, and make $\widehat{P}_j$ more closely conform to P_j.

From (2) and (5), $w_j = \sum_p \phi_p(x_{jp}) - \ln \sum_{k \neq j} e^{v_k}$. Solving for $\phi_q(x_{jq})$ yields

$$(6) \qquad \phi_q(x_{jq}) + \eta_j = - \sum_{p \neq q} \phi_p(x_{jp}) + \ln \sum_{k \neq j} e^{v_k} - \ln \left\{ \frac{1}{\widehat{P}_j(w_j)} - 1 \right\}.$$

using the previous estimate of $\phi_p(\cdot)$, v_k, and $\widehat{P}_j$ in the right hand side.

Now, consider explanatory variable q. For each observation, we have x_{jq} (j=1,2,..,J) and corresponding $\phi_q(x_{jq})+\eta_j$. These form the scatter plot shown in Figure 3.

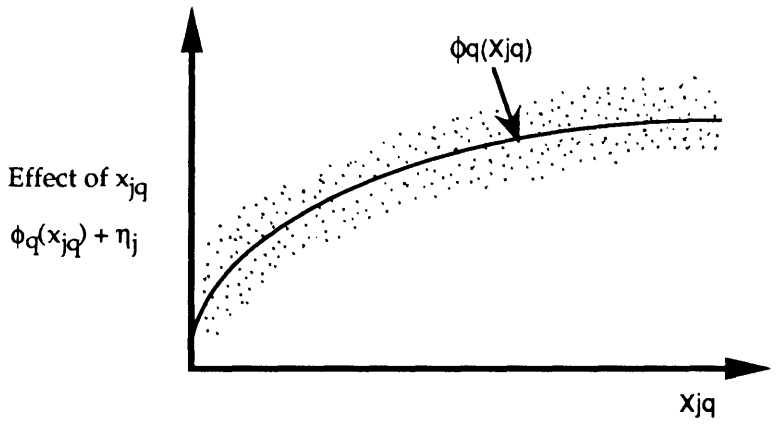


Figure 3: Utility residuals against q-th explanatory variable

A vertical difference between a point and the curve, $\phi q(x_{jq})$, is the utility residual.

A nonparametric regression of $\phi_q(x_{jq})+\eta_j$ on x_{jq} gives the current estimate of $\phi_q(\cdot)$. Iteration over q for several cycles produces the final $\phi_q(\cdot)$'s. In our work, we do all nonparametric regression by kernel estimation (Silverman 1986). Details are discussed in Section 3.3.

2.3 Algorithm

The following algorithm iteratively finds $\phi_q(\cdot)$, where Q denotes the number of explanatory variables, $\Delta\widehat{P}_j(w_j)$ and $\Delta\phi_q(\cdot)$ are square integral difference of $\widehat{P}_j(w_j)$ and $\phi_q(x_{jq})$ respectively between successive iterations.

Initial estimate by linear-in-parameters MNL
Compute initial v_j , w_j's *from the estimated coefficients*
Repeat until $\Delta\widehat{P}_j(w_j) < \delta$

For $q = 1$ *to* Q [loop over explanatory variables]
While $\Delta\phi_q(\cdot) >$ tolerance
Obtain $E(y_j \mid w_j)$ [by nonparametric regression]
Compute $\phi_q(x_{jq})+\eta_j$ [by (6)]
Find $\phi_q(\cdot)$ [by nonparametric regression]
Revise v_j *and* w_j *by MNL using new* $\phi_q(x_{jq})$'s
end [while]
end [for q]
end [repeat]

Convergence of the algorithm is guaranteed simply by adding a statement, "terminate if loglikelihood value decreases" in the repeat loop.

2.4 Summary: Two key concepts of URM

A crucial idea of URM is the use of the utility residuals, which are latent and unobservable in a discrete choice case. In order to find them, we postulate the logit distributional assumption Thus we are dealing with a class of semiparametric methods in which model structure is nonparametric and distributional assumptions are parametrically specified. This is in contrast to traditional semiparametric literature in econometrics, so called distribution-free methods, where minimum assumptions are imposed on the disturbance distribution while model specification is parametric. These include the maximum score estimator (Manski, 1975, 1986, 1989), the maximum likelihood estimator (Cosslett 1983), the quasi-maximum likelihood

estimator (Klein & Spady 1988), and regression models of Duncan (1986), Stoker (1986), and Han (1987).

The author is aware of one working paper in which Matzkin (1989) pursues an idea of the specification-free semiparametric approach in general multinomial models using a constrained maximum likelihood estimation to obtain a multidimensional nonparametric utility function. In the field of statistics, generalized additive models (Hastie and Tibshirani 1986, 1987) pursue nonparametric additive model specification with a parametric link function for distributional form. Their binary logistic regression case does not directly apply to MNL because doing so violates its fundamental assumption of the additivity. Their extension to a matched case-control model, however, can be readily used in MNL with slight notational modifications and possesses many attractive theories associated with General Additive Models (GAM). Although its application to the simulation data presented in Section 3 leads to the similar result as URM, the extended GAM is not quite robust and has failed on several real databases. Appendix C discusses GAM and its extension to MNL, and presents both simulation study analogous to the one for URM in Section 3 and a real data application which has failed .

3. SIMULATION STUDY

3.1 First Cut Result

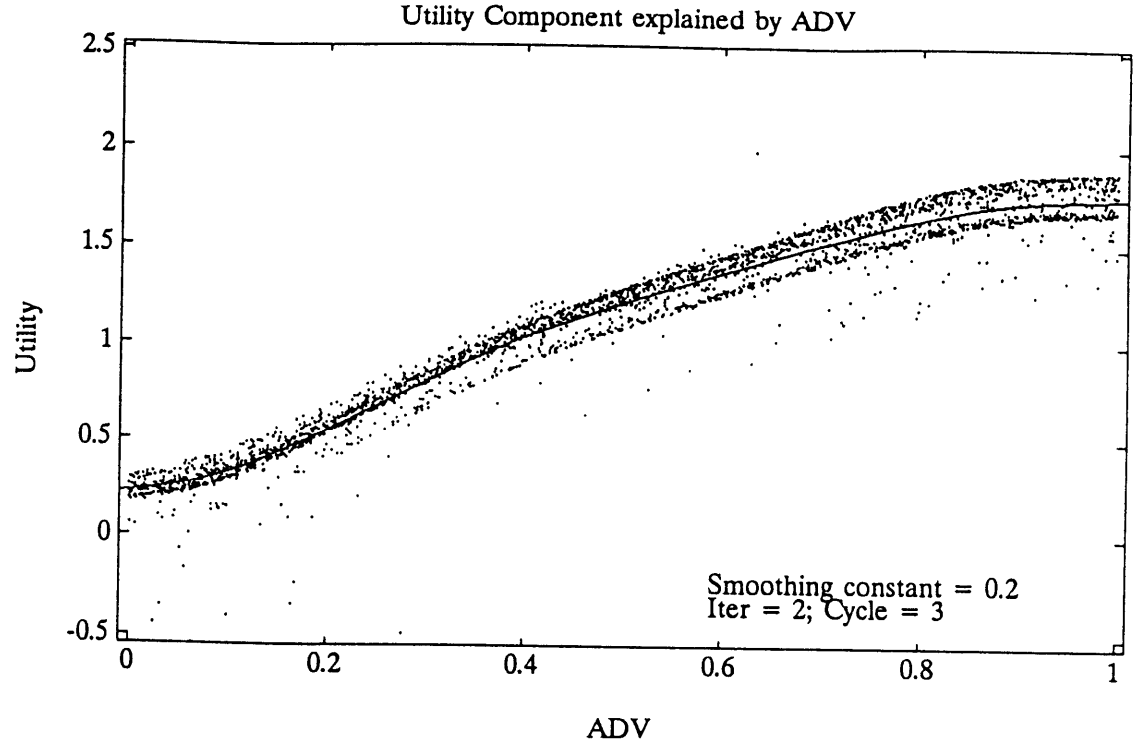
To test and evaluate the validity of URM, first, a simulation study is conducted from a known model. Multinomial choice data is generated from a pre-specified nonlinear additive separable utility function, which is then recovered by URM. The data contains three alternatives with 988 observations. Marketing mix variables are based on an actual Aseptic Drink database. However, in this database, price occurs at discrete values such that the last digit of the cents is either 4 or 9 in the price range of 55~99 cents (see Figure 17 of utility v.s. price plot in Section 4). This is an undesirable distraction for testing URM. Therefore, the simulated price is randomly generated from the uniform distribution between 50 cents to $1 for each alternative. Because the database does not contain advertising exposure information, this is also randomly generated from the uniform distribution between 0 and 1. The systematic component of the utility function for alternative j is then pre-specified as follows.

$$v_j = 0.317 \cdot asc2_j + 0.307 \cdot asc3_j + 0.567 \cdot feature_j$$

$$+ 0.700 \cdot display_j + \log (adv_j+0.1) - 2 \cdot (price_j)^3$$

asc2 and asc3 are alternative specific constants, feature and display are binary promotional activities whose values are taken from the data. The coefficients are selected based on estimates from previous linear logit model runs in order to have realistic orders of magnitude. Note that a diminishing return effect of advertising is assumed by means of the logarithmic function, and price exhibits increasing marginal disutility expressed by the negative cubic term.

Figure 4 shows the resulting nonparametric plots of $\phi_{adv}(\cdot)$ and $\phi_{price}(\cdot)$ for each variable after convergence is achieved with 3 cycles of the outer loop. The absolute level of the utility scale is irrelevant and the entire curve can be shifted to have the mean of zero, since multinomial logit takes into account of only the differences among alternatives. Both exhibit an excellent recovery of the true underlying nonlinear relationships, the logarithmic for ad exposure and negative cubic for price. The top graph in Figure 5 is an empirical probability plot generated from the utility residuals right after the initial linear MNL run (Cycle 0) and the bottom one is at the end of the three URM cycles (Cycle 3). Notice how closely the final residuals are distributed to the theoretical logistic function. The deviation between the empirical and theoretical probability plot at Cycle 0 is solely due to the misspecification by modeling the utility function as a linear form other than sample variations.



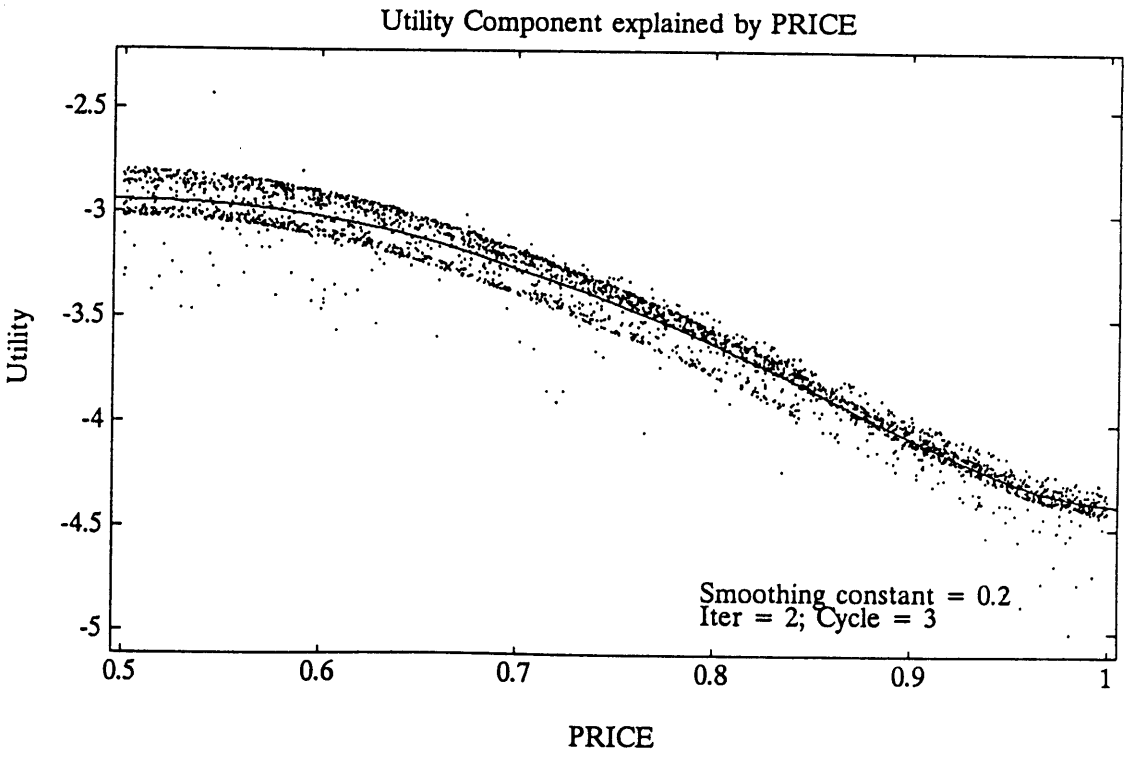
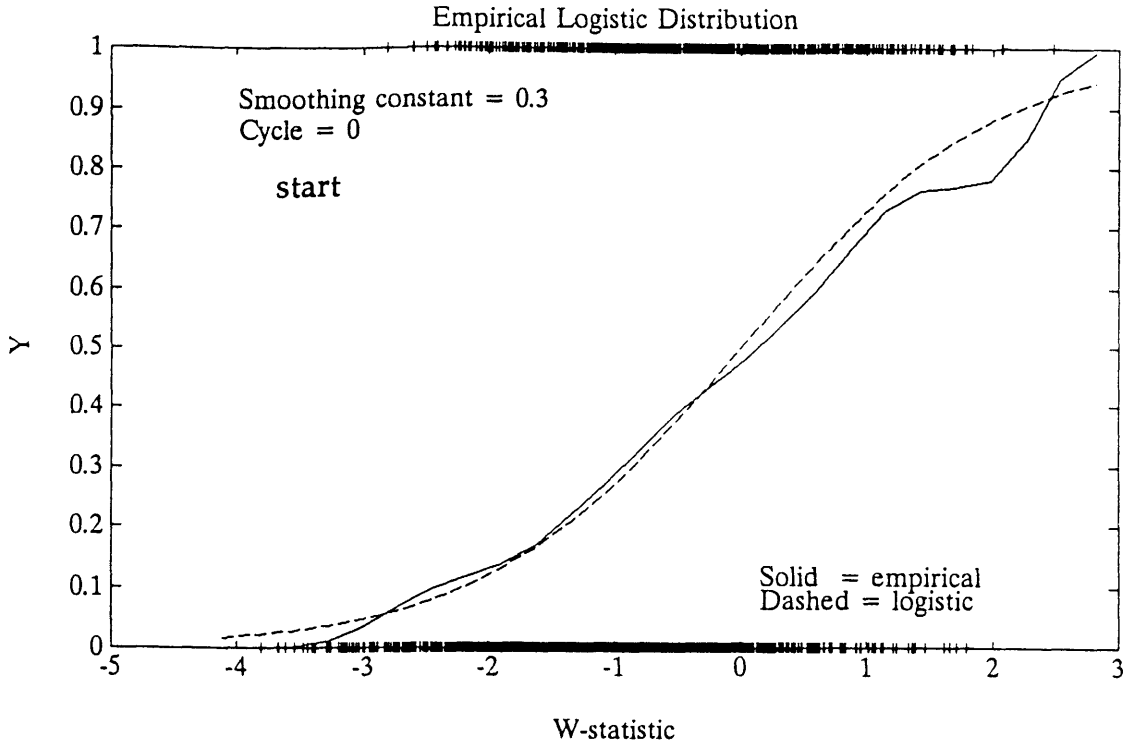


Figure 4: Additive nonparametric utility transformations by URM in the simulation study



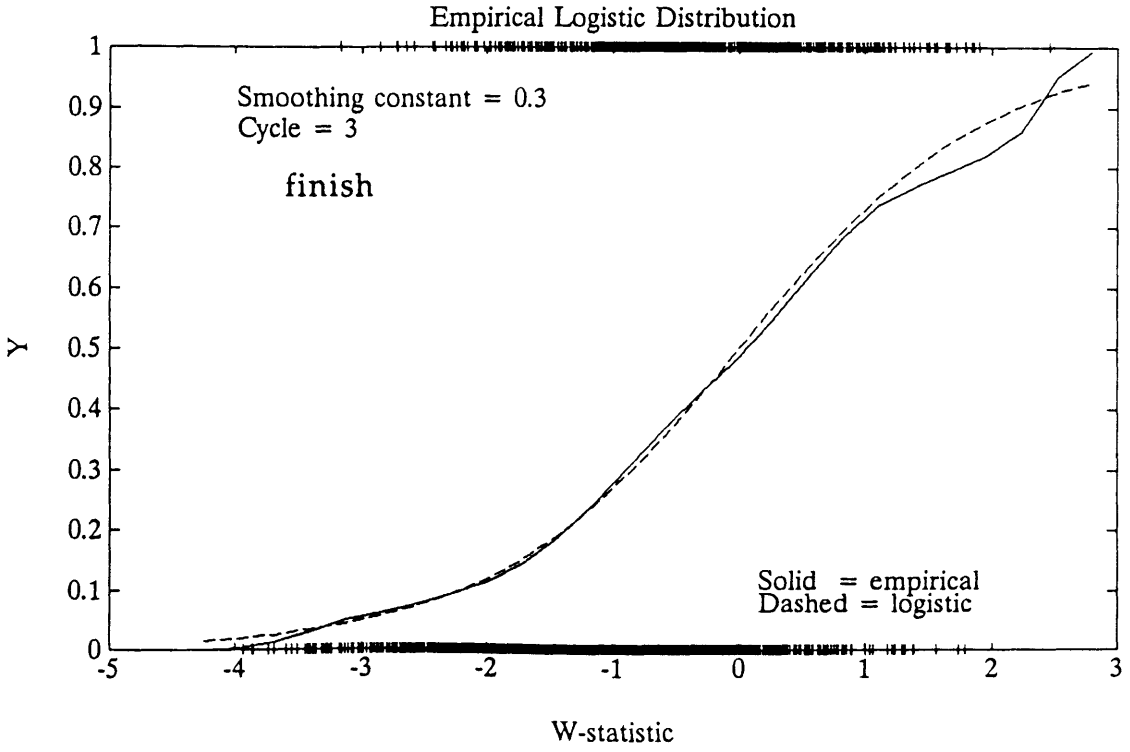


Figure 5: An empirical and theoretical distribution generated from utility residuals in the simulation study

Table 1: Result of linear and the URM logit in the simulation study

Linear Model

variable	coeff.	std.err	t-stat
FEATURE	0.5811	0.1189	4.8850
DISPLAY	0.7681	0.2142	3.5863
ADV	1.8391	0.1585	11.6049
PRICE	-3.4936	0.3189	-10.9558
ASC2	0.2098	0.0932	2.2516
ASC3	0.2860	0.0918	3.1167

mean probability of correct choice = 0.4379 $\rho^2 = 0.14643$
percentage of correct choice = 54.45 loglikelihood value = -926.49

mean absolute deviation = 0.0904 mean absolute second derivative = 0.4693

The URM Logit Model

variable	coeff.	std.err	t-stat
FEATURE	0.5844	0.1197	4.8816
DISPLAY	0.8102	0.2161	3.7484
ϕ(ADV)	1.1105	0.0929	11.9581
ϕ(PRICE)	1.2180	0.1069	11.3923
ASC2	0.2108	0.0939	2.2461
ASC3	0.2934	0.0923	3.1797

mean probability of correct choice = 0.4466 $\rho^2 = 0.15987$
percentage of correct choice = 54.66 loglikelihood value = -911.90

mean absolute deviation = 0.0210 mean absolute second derivative = 0.3172

The True Underlying Model

variable	coeff.	std.err	t-stat
FEATURE	0.5883	0.1199	4.9083
DISPLAY	0.8036	0.2157	3.7246
log(ADV+0.1)	1.0108	0.0848	11.9263
$PRICE^3$	-2.1616	0.1915	-11.2883
ASC2	0.2141	0.0938	2.2813
ASC3	0.2957	0.0923	3.2026

mean probability of correct choice = 0.4469 $\rho^2 = 0.16057$
percentage of correct choice = 54.05 loglikelihood value = -911.14

Table 1 illustrates how well URM recovered the original function by applying linear, URM, and the true model transformation of the advertising exposure and price for MNL estimation. An examination of U-square, a measure of loglikelihood, suggests that URM captures a striking 92% of the information gained by going from the linear null model to the correct model specification. (Hauser 1978) A bootstrap estimate using 50 samples gave 89.4% recovery in the log likelihood values with an asymptotic standard error of 3.4%. Furthermore, similar magnitudes of improvement (over 90%) have been observed in replications of the simulation runs using four other sets of random numbers as well. The mean absolute deviation figures in the table show better correspondence between the empirical and theoretical probability plot after URM. The mean absolute second derivative is used to monitor its smoothness.

Another finding from the table is a surprisingly good prediction by the linear model. In fact, the linear model surpasses the true model in the percentage of correct choice, and only a slight degradation can be seen in the probability of correct choice. In other words, as far as prediction is concerned, a linear utility MNL model performs quite well even if it is misspecified. Once again, this confirms the robustness of the linear-in-parameter MNL in cross-validations, the fact demonstrated by many empirical studies.

The computational time varies from run to run depending on how many cycles and inner iterations are repeated. In all cases, convergence is achieved within three cycles and three iterations, and actual time is between 5 to 8 minutes on a 25 MHz 486 machine.

3.2 Robustness against Various Distributions for the Disturbance Term

3.2.1 Testing the Distributional Assumption of URM

The fundamental idea behind URM is that any discrepancy between the theoretical and empirical probability plot is caused by model misspecification. However, this reasoning is supported only if the discrepancy is not largely induced by incorrect assumption of the disturbance distribution (i.e. doubly exponential) in the stochastic utility function. We test this by using a disturbance term with different distributions.

Five sets of the simulated choice data are generated from the same utility specification as before but with different disturbance distributions; doubly exponential, uniform, independent normal, normal with correlation of 0.5 between alternative 1 and 2, and normal with correlation of 0.8 between alternative 1 and 2. They are all derived from the same set of random numbers

(using the same seed) to minimize instability, and their variances are set to $\pi^2/6$ for scale consistency so as to allow a direct comparison of the magnitude of the estimators.

Figures 6 and 7 show the estimated advertising and price utility transformations respectively. While comparable curves are expected for the independent normal since it resembles the doubly exponential, we find that the uniform distribution also results in the similar transformations. The conclusion drawn from these three plots is that URM is quite robust against its distributional assumption in recovering the underlying utility structure.

3.2.2 Violation of IIA

The normal distributions with correlation involve a tougher test, not only of the URM assumptions but also simultaneously MNL on violation of independence among alternatives. In spite of this, the transformations approximate the correctly assumed curve fairly well as long as the correlation is within a reasonable level (0.5). For the excessively high correlation of 0.8, some deviation is apparent although the general shape is still preserved. The high correlation has much stronger impact on estimation bias in other variable coefficients. The estimates for the 0.8 correlation case are 0.820 (t=6.60) for feature, 1.145 (t=5.12) for display, 0.272 (t=2.72) for asc2, and 0.643 (t=6.71) for asc3, in comparison with 0.640 (t=5.42), 0.839 (t=3.92), 0.101 (t=1.10), and 0.195 (t=2.15) respectively for the uncorrelated normal distribution. The bias seems to be inherent in a correlated disturbance. Such bias is observed in the normal with 0.5 correlation to a lesser degree but not in any of the uncorrelated distributions. This study empirically supports the notion that URM is quite stable under wide variety of distributional assumptions, helping to justify the logic behind URM as well as to extend its applicability in various real situations.

3.3 Issues in Nonparametric Regressions

Within each inner loop of URM algorithm, two types of nonparametric regression are executed for different purposes; one regresses a binary choice indicator y_j on w_j to compute the empirical probability plot, $E(y_j | w_j)$, and the other regresses the partial utility residuals on marketing mix variable x_q to obtain nonparametric utility transformation, $\phi_q(x_q)$. Because the algorithm does not explicitly specify the nature of these nonparametric regressions, they will be investigated in the following subsections.

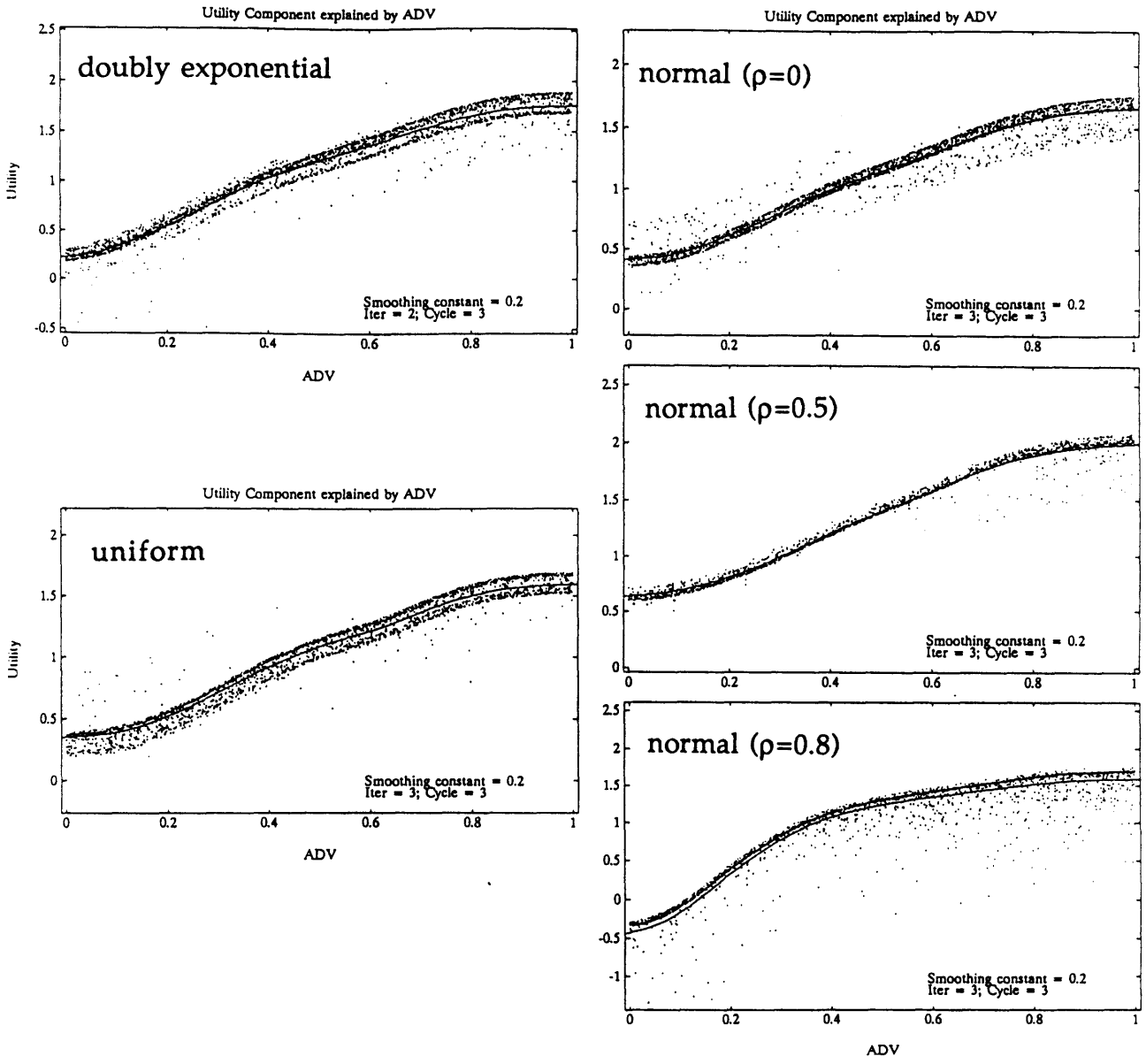


Figure 6: Advertising utility transformation with different error distributions

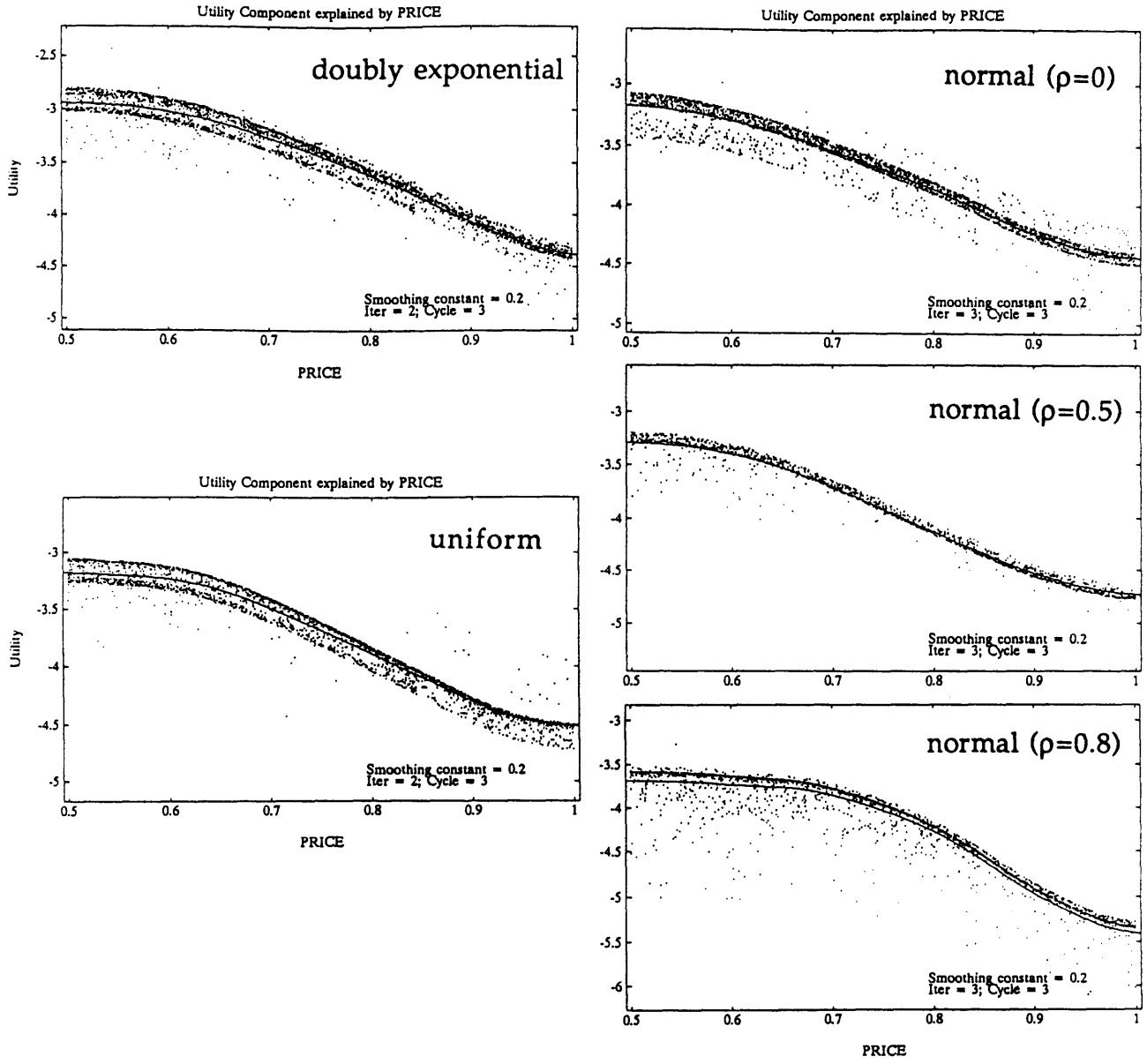


Figure 7: Price utility transformation with different error distributions

3.3.1 Existing methods in nonparametric regression

Since both nonparametric regressions involve only a single explanatory variable, one can choose from several well-established techniques as well as some simple smoothing schemes. Examples of the former are kernel regression (Nadaraya 1970, Watson 1964) and various splines (Wegman and Wright 1983), while running mean, running line, and running median belong to the latter. See Hastie and Tibshirani (1989) for an overview of each technique as a smoothing device. A nonparametric regression technique for URM should be selected based on the following three criteria.

First, the resulting fit must produce a smooth curve. We can reasonably expect both the probability plot and the utility transformation functions to be continuous since they model human behavior. Running mean, running line, and running median are not suitable in this regard because of discontinuity in the first derivative at each observation.

Second, it is desirable if the fit of the regression $\hat{\mathbf{y}}$ can be expressed as linear in observed response values y_i's. That is, $\hat{\mathbf{y}} = S\,\mathbf{y}$, where $\hat{\mathbf{y}}$ and $\mathbf{y}$ are $n\times 1$ column vectors, S is a $n\times n$ matrix, and n is the number of observations. In this case, many useful concepts and properties, such as degrees of freedom and standard error, can be derived by the analogy of linear regression. Running mean, running line, kernel, and splines are all linear operations, while running median and variable kernel whose smoothing depends on the values of $\mathbf{y}$ are not.

Finally, but not least, the computational efficiency is another important criterion since the regression routine is repeated many times in the URM algorithm.

In addition to these considerations, ease for extending to higher dimensions leads us to adapt kernel regression, which is expressed as

$$(7) \qquad E(y \mid x) = \frac{\sum_i y_i K\left(\frac{x - x_i}{h}\right)}{\sum_j K\left(\frac{x - x_j}{h}\right)} .$$

where $K(\cdot)$ is called a kernel function. We used the popular Gaussian function. h is a smoothing constant which essentially determines the effective distance of the i-th observation to its neighbors. It is well known that the resulting estimate is much more sensitive to the

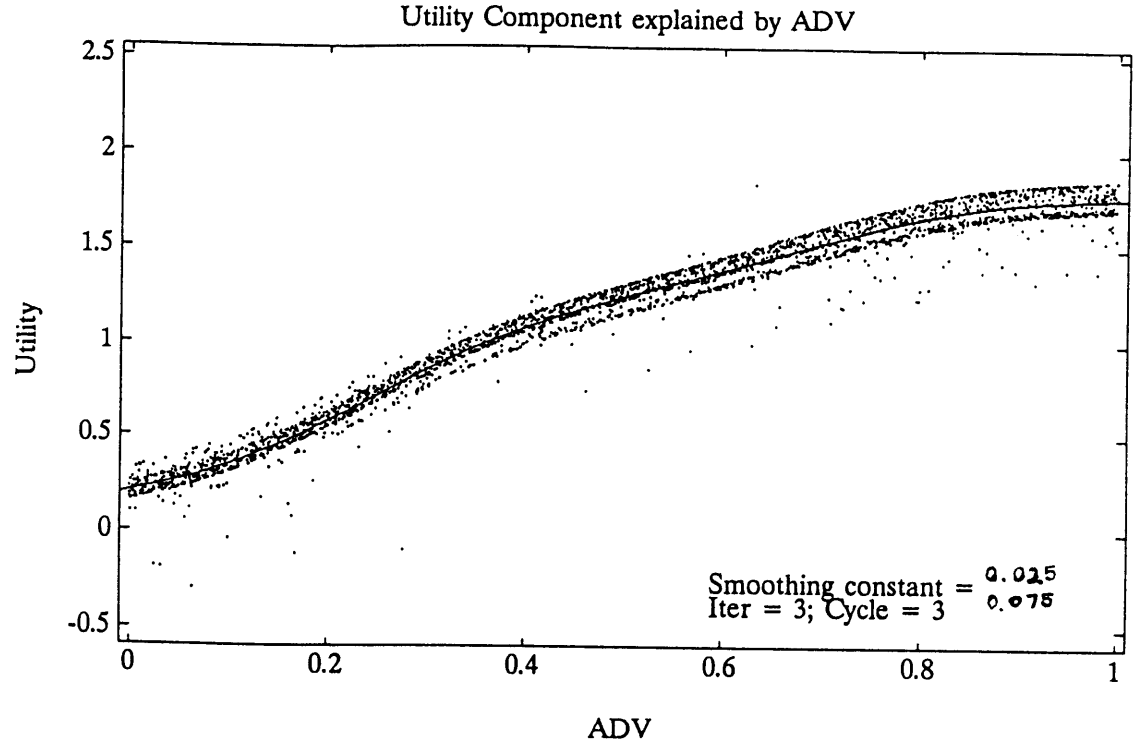
choice of the smoothing constant than the functional form of the kernel. (Silverman 1986, Ullah 1988) A detail discussion on its choice appears in subsection 3.3.4. In short, $E(y \mid x)$, can be interpreted as a local weighted average of all observations y_i's, in which the weight of each point is determined by its distance from x.

The kernel regression has two minor drawbacks, which are also shared by some other nonparametric regression techniques. First, it exhibits biases at the boundaries as seen in Figure 4. This is due to the effect of the weighting from only one side. The problem can be partly alleviated by using a kernel function which puts more weight along the line of fit. A moving ellipsoid method proposed by Abe(1991) adaptively controls the shape of the kernel function in such a manner. Figure 8 is the final utility transformations obtained by the moving ellipsoid method. Although some improvements are observed, its marginal effect does not seem to be large in our class of applications, and so our work here will use the standard kernel regression.

A second weakness is that the estimator is sensitive to outliers. While it does not appear to pose any problem in the simulation study, it could potentially become relevant in actual data. One simple solution is to preprocess data by applying a running median of three before the kernel regression is run. I will refer to this as robust kernel regression in contrast to the standard kernel regression which is not robust. Because outliers can possibly contain important information, whenever the robust regression is run it is always advisable to compare with the non-robust version as well. Figure 9 shows the final transformations using the robust kernel regression. As one can see, partial residuals are clustered more tightly along the curves compared with the non-robust result. Note that neither the moving ellipsoid method nor running median are linear operations.

3.3.2 Degrees of freedom

To facilitate various inferences, it is very useful to have a notion of degrees of freedom. While nonparametric regression does not contain parameters to be estimated, one could devise a concept of equivalent degrees of freedom which are mainly controlled by the smoothing constant. The intuition is as follows. As h goes to infinity, the kernel function looses its local weighting property and approaches the global averaging of all observations. In this case, the degrees of freedom is one because it is equivalent to having just an intercept term in linear regression. As h is reduced slightly, the fit might resemble a near straight line or smooth unimodal curve which is comparable to linear or quadratic approximation. This should correspond to low degrees of



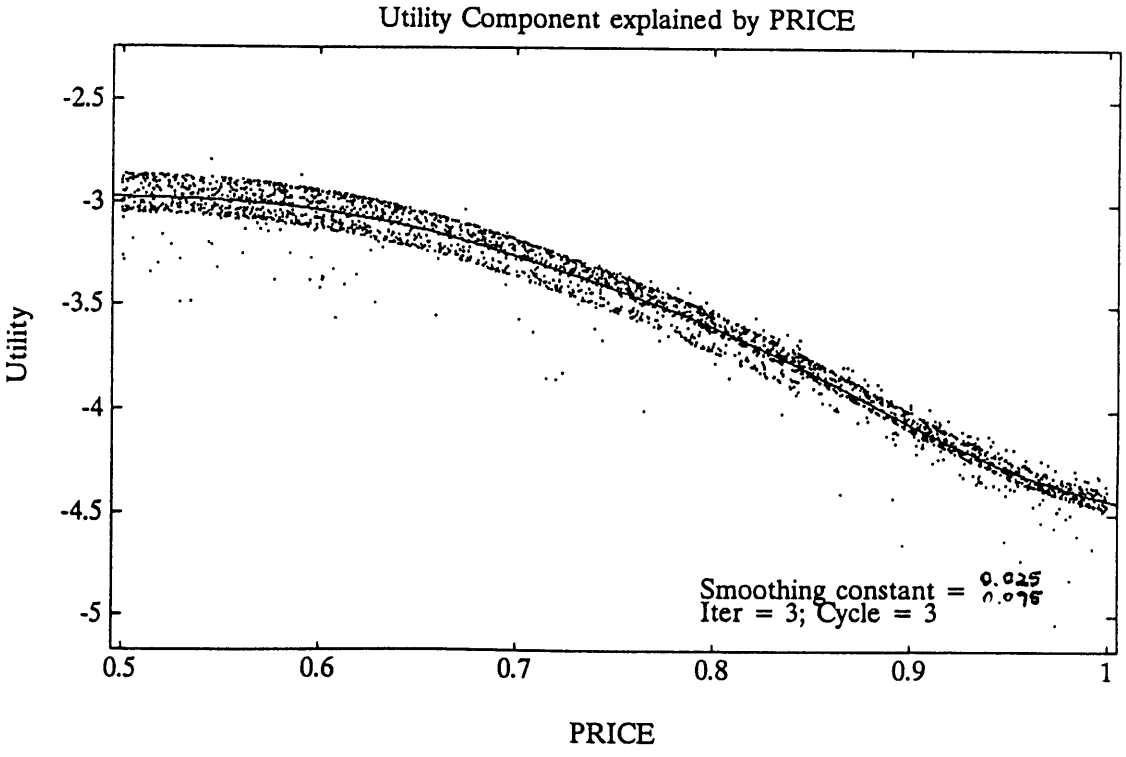
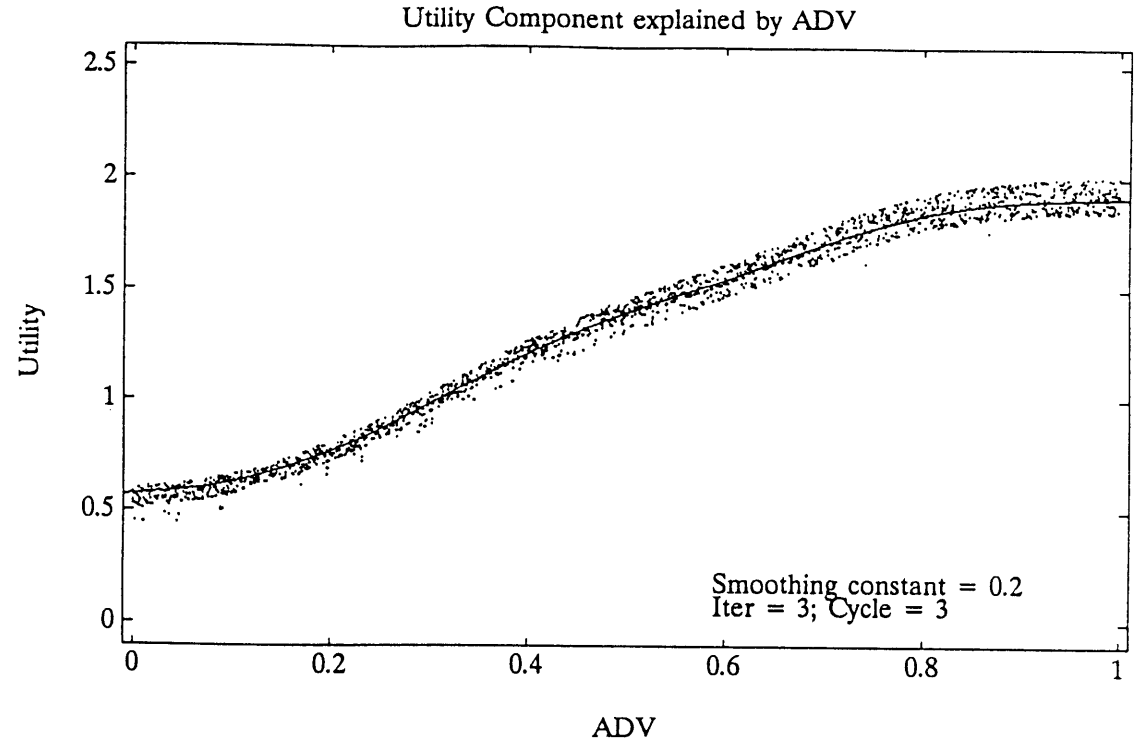


Figure 8: Additive nonparametric utility transformations using the moving ellipsoid method



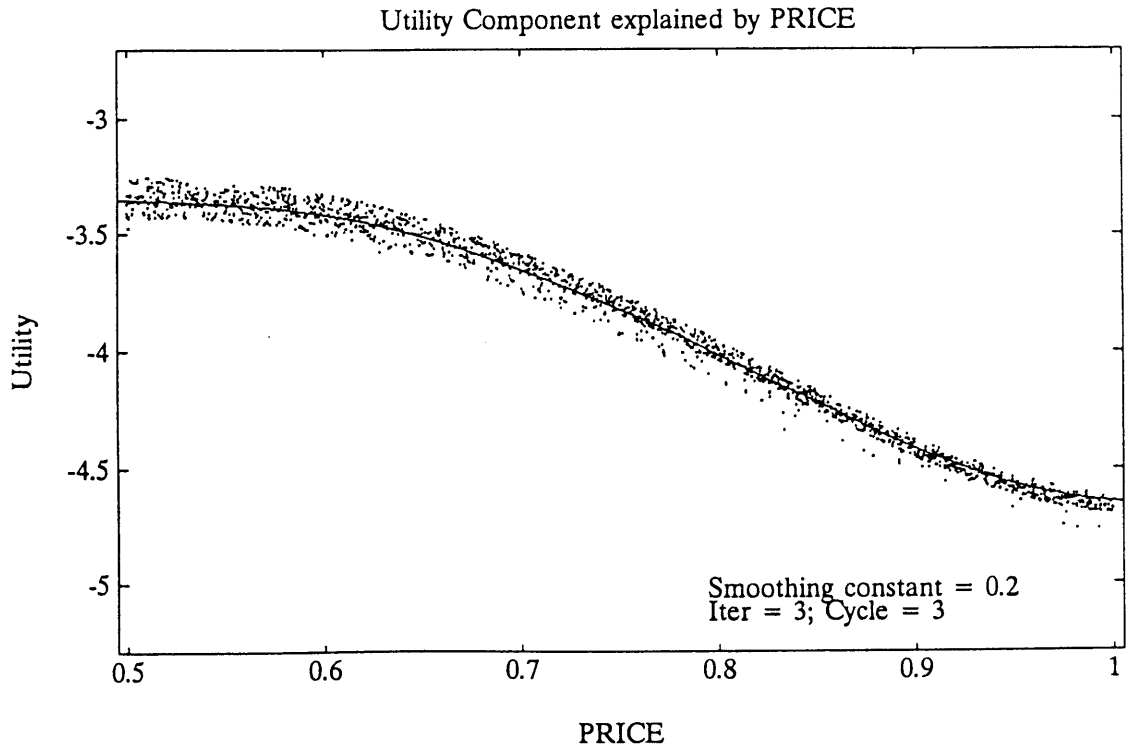


Figure 9: Additive nonparametric utility transformations using robust kernel regression

freedom somewhere between 2 and 3. When h is much smaller, the curve becomes quite wiggly, which leads to high degrees of freedom.

For linear nonparametric estimator, $\hat{\mathbf{y}} = \mathbf{S}\,\mathbf{y}$, one can formulate the degrees of freedom in analogy to linear regression where the estimator can be expressed as

$$\hat{\mathbf{y}} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{y} = \mathbf{H}\,\mathbf{y}\,.$$

H is referred to as a hat matrix which is a projection matrix. Three formulations for the degrees of freedom suggested by Hastie and Tibshirani (1989) are briefly reviewed below. Here, the model is assumed to have a form, $\mathbf{y} = \mathbf{f} + \varepsilon$, with ε being i.i.d. with mean 0 and variance σ^2, y is a n×1 observation vector, and f is a linear regression model with p parameters.

[1] tr(SS')

Sum of variances for prediction can be expressed as

$$\begin{aligned}\sum_i \mathrm{Var}(\hat{y}_i) &= \mathrm{tr}\left(\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\\ &= \mathrm{tr}\left(\mathbf{H}\,\mathbf{y}\,\mathbf{y}'\mathbf{H}'\right)\\ &= \mathrm{tr}\left(\mathbf{H}\,\mathbf{H}'\right)\sigma^2\\ &= p\,\sigma^2\end{aligned}$$

[2] tr(2S - S'S)

Expected residual sum of squares is

$$\begin{aligned}\mathrm{ERSS} &= \mathrm{E}(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})\\ &= \left[\,n - \mathrm{tr}\left(2\mathbf{H} - \mathbf{H}\,\mathbf{H}'\right)\right]\sigma^2 + \mathbf{b}'\,\mathbf{b}\end{aligned}$$

where $\mathbf{b} = \hat{\mathbf{f}} - \mathbf{f}$ is a bias term.

[3] tr(S)

Correction factor to be added to the residual sum of squares to obtain C_p statistic (Mallows 1973, Meyers 1986) is $2\,\mathrm{tr}(\mathbf{H})\,\sigma^2$.

In the case of linear regression, all three produce the identical degrees of freedom, p, since H is a projection matrix. If S is a symmetric matrix whose eigenvalues are between 0 and 1, it is easy to show that $tr(S\,S') \le tr(S) \le tr(2S - S'S)$. Because tr(S) is between the other two in such a case and is the simplest to compute, this formula will be used as the equivalent degrees of freedom from now on. It also has an intuitive appeal. For one side of extreme cases, if S is to simply take the global average, S is a diagonal matrix whose elements are all equal to 1/n and the degrees of freedom becomes 1. On the other hand, if no smoothing takes place at all and hence the fit simply interpolates the observed points, S is an identity matrix and its degrees of freedom corresponds to n, the number of observations.

When the number of observations is large, even a computation of tr(S) becomes time consuming. A good approximation formula is

$$(8) \qquad tr(S) = \frac{1}{\sqrt{2\pi}} \frac{1}{h} \frac{\sigma_x}{\sqrt{1/12}} \alpha\left(\frac{h\sqrt{1/12}}{\sigma_x}\right)$$

where $\alpha(\cdot)$ is given in Table B of Appendix B. For the derivation, see the appendix.

Table 2 is a result of goodness-of-fit for various nested models using the notion of the degrees of freedom, (8). Deviance is defined as -2 times loglikelihood to facilitate the chi-square test. The full URM model with price and advertising nonparametric transformations is better than the full linear utility model at 1% level of significance. But when either price or advertising variable is missing, URM is not significantly better than its linear counterpart. This implies the importance of including all necessary predictor variables in order to gain the maximum performance out of URM.

Table 2: Goodness-of-fit of Various Nested Models

model	deviance	Δdeviance*	df	Δdf*
asc2, asc3	2169.10		2	
asc2, asc3, feature, display	2129.18		4	
asc2, asc3, feature, display, price	2002.50		5	
asc2, asc3, feature, display, ϕ_{price}	1996.34	6.16	11.4	6.4
asc2, asc3, feature, display, adv	1985.02		5	
asc2, asc3, feature, display, ϕ_{adv}	1975.08	9.94	11.4	6.4
asc2, asc3, feature, display, price, adv	1852.98		6	
asc2, asc3, feature, display, ϕ_{price}, ϕ_{adv}	1824.88	28.10**	18.8	12.8**

*** Δdeviance and Δdf refers to the difference between the URM and its linear counterpart (line above).**

**** Significant at 1%**

3.3.3 Pointwise standard error

Pointwise standard error is calculated from $\text{cov}(\hat{\mathbf{f}}) = S S' \sigma^2$. Assuming that the bias, $E(\hat{\mathbf{f}}) - \mathbf{f}$, is negligible (which is very difficult to check), 2 times square root of the diagonal elements can be used to obtain an approximate 95% confidence band. In practice, however, the repeated applications of nonparametric regression within both inner and outer loop makes the explicit evaluation of the final S almost infeasible. Therefore, we have turned to bootstrap estimates (Efron 1981, Efron and Gong 1983), which is described below.

Let θ_b be an estimate of parameter θ based on random sample of size n with replacement from the observed data. Repeat this process B times. Then, the variance estimate of θ is

$$\text{Var}(\theta) = \frac{1}{B-1} \sum_{i=1}^{B} (\theta_b - \bar{\theta})^2$$

where $\bar{\theta}$ is a mean of θ_b.

Figure 10 shows the result of 50 bootstrap estimates for advertising and price utility transformations. Based on them, 95% pointwise confidence bands are computed in Figure 11. The bands are narrow enough to support the nonlinearity of the transformations.

3.3.4 Smoothing constant

As equation (7) shows, the smoothing constant, h, determines how fast an influence of a particular observation point decays as the distance. As shown in Part I, h actually controls the trade off between bias and variance of the estimated regression curve. For clarity, let us distinguish the smoothing constant h associated with the binary regression for the empirical probability curve and utility transformations for advertising and price by adding subscript as h_b, h_a, and h_p, respectively when necessary.

A standard automatic method for choosing the value of h is a cross validation. However, in URM we rely on a graphical and subjective approach for the following reasons.

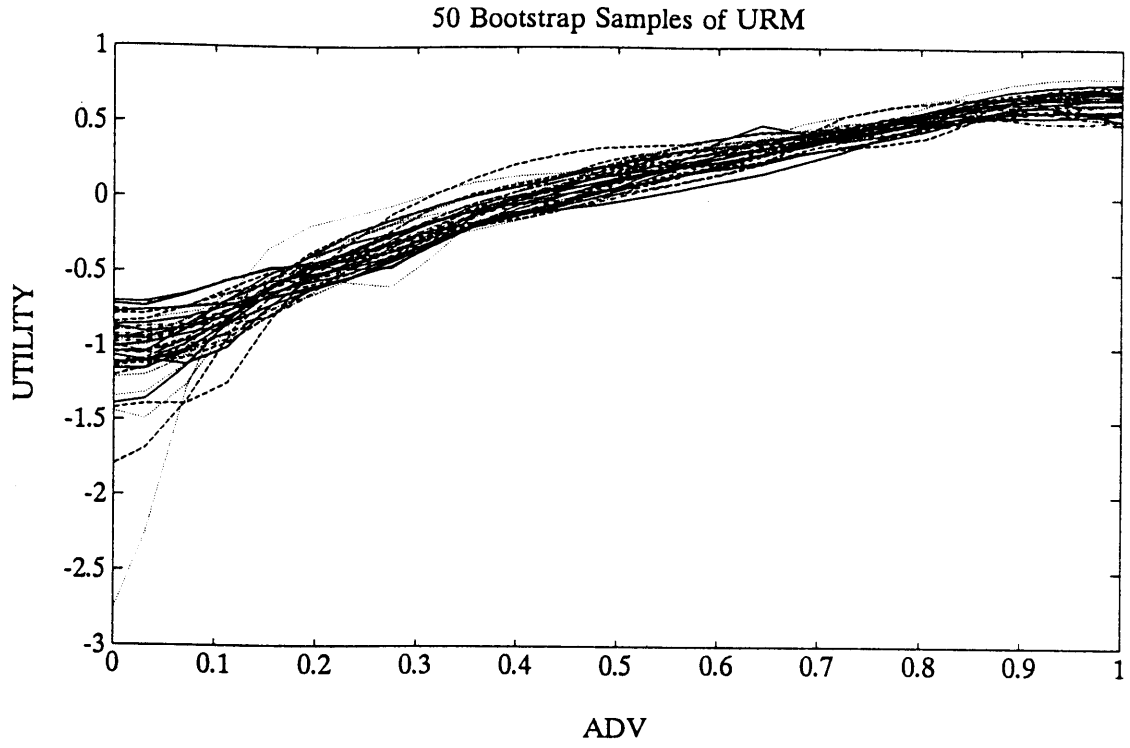
[1] The computational requirements of cross validation for a nonlinear model such as multinomial logit would be extremely high.

[2] The performance of cross validation in practice has not been very good. (Hastie and Tibshirani 1990)

[3] To some extent, the best choice of h in terms of usefulness of the final result depends on the user's prior belief. In a one dimensional case such as URM, a managerial judgement can be easily made on the basis of visual inspection.

The third reasoning especially applies to the nonparametric regression for the utility transformations. For instance, in the simulation study, an analyst should suspect undersmoothing in the adv (price) transformation if the curve is not monotonically increasing (decreasing).

On the other hand, care must be exercised for the other nonparametric regression which computes the empirical probability plot, $E(y \mid w)$, whose deviation from a logistic function directly affects the magnitude of the utility residuals. It could be the case that a small change in h_b greatly influences the size of the residuals and consequently the shape of the transformations and overall fit of the model. Furthermore, managers in general have little prior knowledge and expectation on how the probability plot should look for a given utility



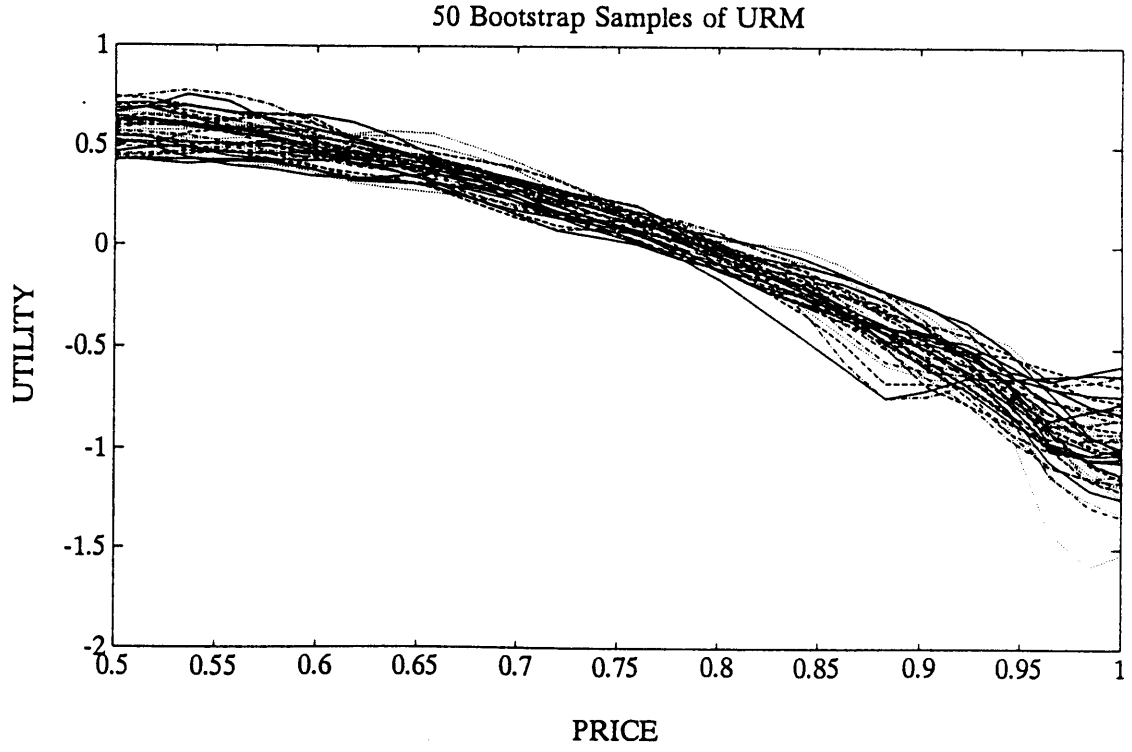
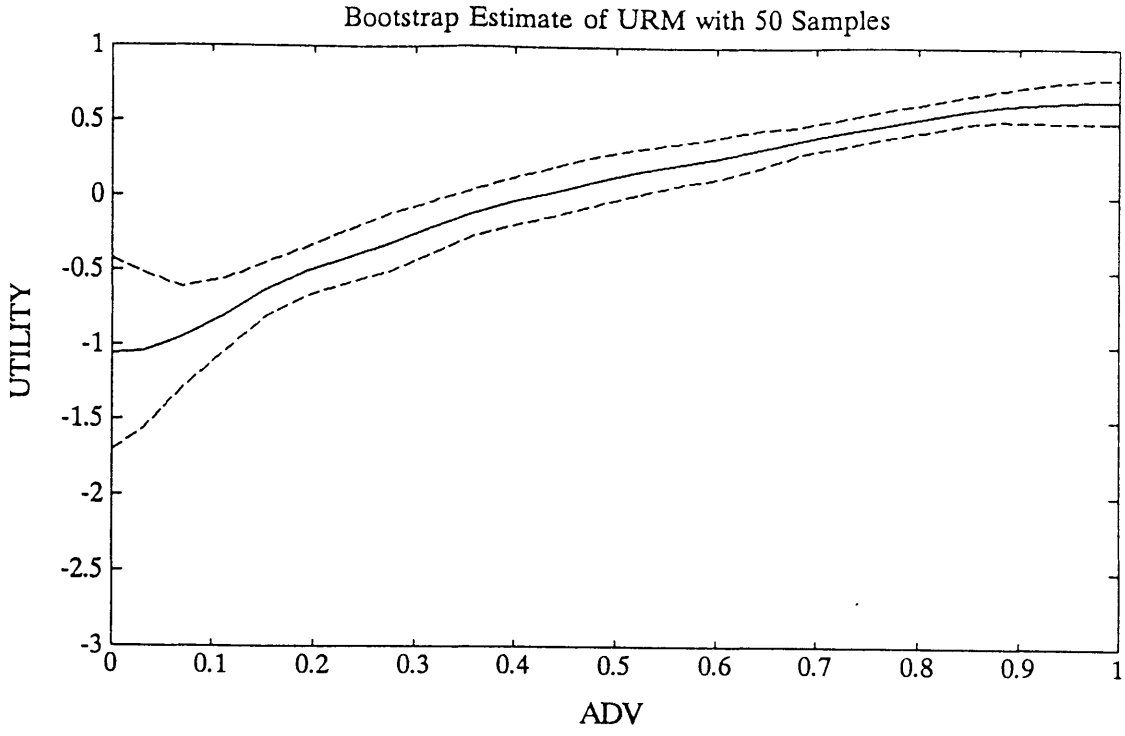


Figure 10: Bootstrap estimates of utility transformations with 50 samples



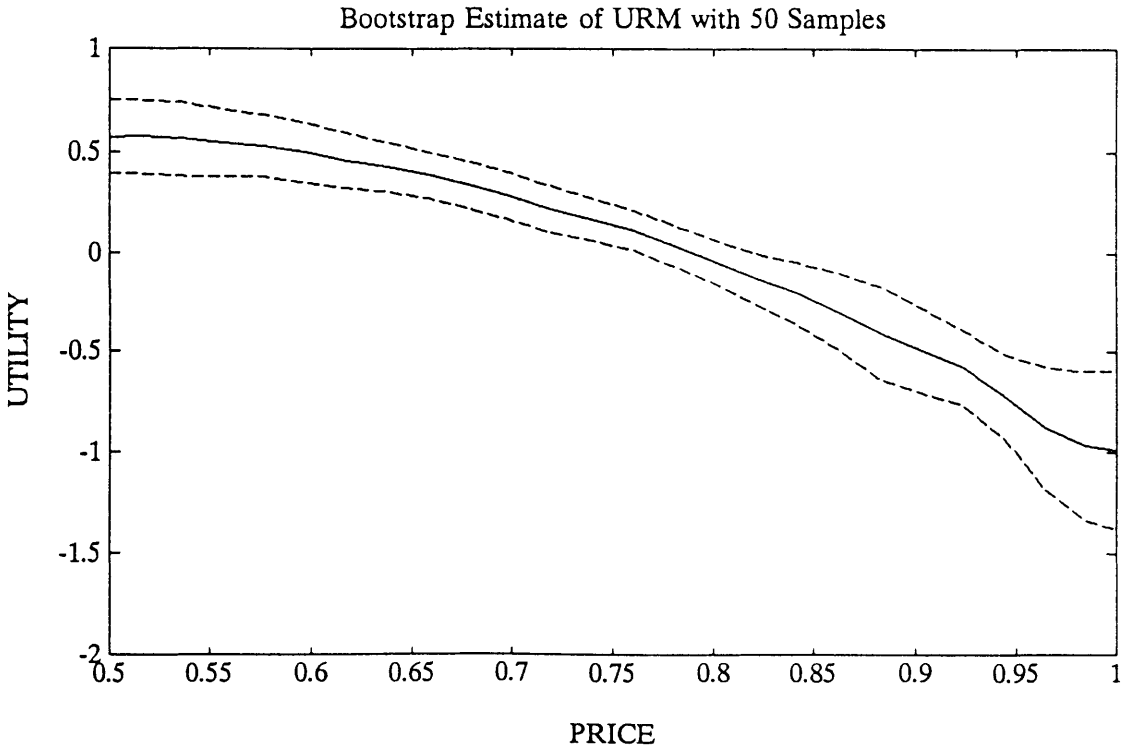


Figure 11: Approximate 95% Confidence band for the utility transformations based on 50 bootstrap estimations

specification. Therefore, a sensitivity analysis is conducted by evaluating URM at seven different values of the smoothing constant for the binary regression, E(y | w).

Figure 12 shows the probability plots in the final iteration, and Figures 13 and 14 are the resulting advertising and price utility transformations respectively. The general shape is preserved rather well except for extreme values of h_b. Undersmoothing can be easily detected if the probability plot is double valued as in the case of h_b=0.05 and 0.10. The rule of thumb is to choose the smallest value of h_b which avoids the double value.

There is also a more systematic way for determining the smoothing constant h_b. Table 3 exhibits ρ^2, deviance, and approximate degrees of freedom using formula (8). In order to investigate whether a decrease in the smoothing constant improves the fit significantly considering that the degrees of freedom increases, the chi-square test is conducted for each value of h_b against the null hypothesis which corresponds to the value of h_b in the line below. This infers the optimal value of h_b by the analogy of a test for nested models in parametric models. The result shows that a model for h_b=0.30 with df=5.1 cannot be rejected at 1% level of significance when the null hypothesis of a more restricted model of h_b=0.40 with df=3.8. Similarly, a model with df=7.4 for h=0.20 cannot be rejected at 5% when the null model of h_b=0.30 has df=5.1. Although this method is an approximation, it is quite adequate considering the fact that the transformations are not very sensitive to the choice of h_b as demonstrated in Figures 13 and 14. In terms of computation, it requires only 7 runs of URM, which is substantially less than cross validation.

Table 3: Goodness-of-fit for different values of h_b in the binary kernel regression

h_b	ρ^2	df	Δdf*	deviance	Δdeviance*	χ^2-test v.s. next below
0.05	0.15451	27.8	13.6	1862.8	-16.4	not significant
0.10	0.15789	14.2	6.8	1846.4	2.8	not significant
0.20	0.15987	7.4	2.3	1849.2	8.2	sig. at 5%
0.30	0.16013	5.1	1.3	1857.4	10.4	sig. at 1%
0.40	0.16046	3.8	0.6	1867.8	12.6	sig. at 1%
0.50	0.16015	3.2	0.4	1880.4	14.8	sig. at 1%
0.60	0.16005	2.8	n.a	1895.2	n.a	n.a

*** Δdf and Δdeviance refers to the difference from the line below**

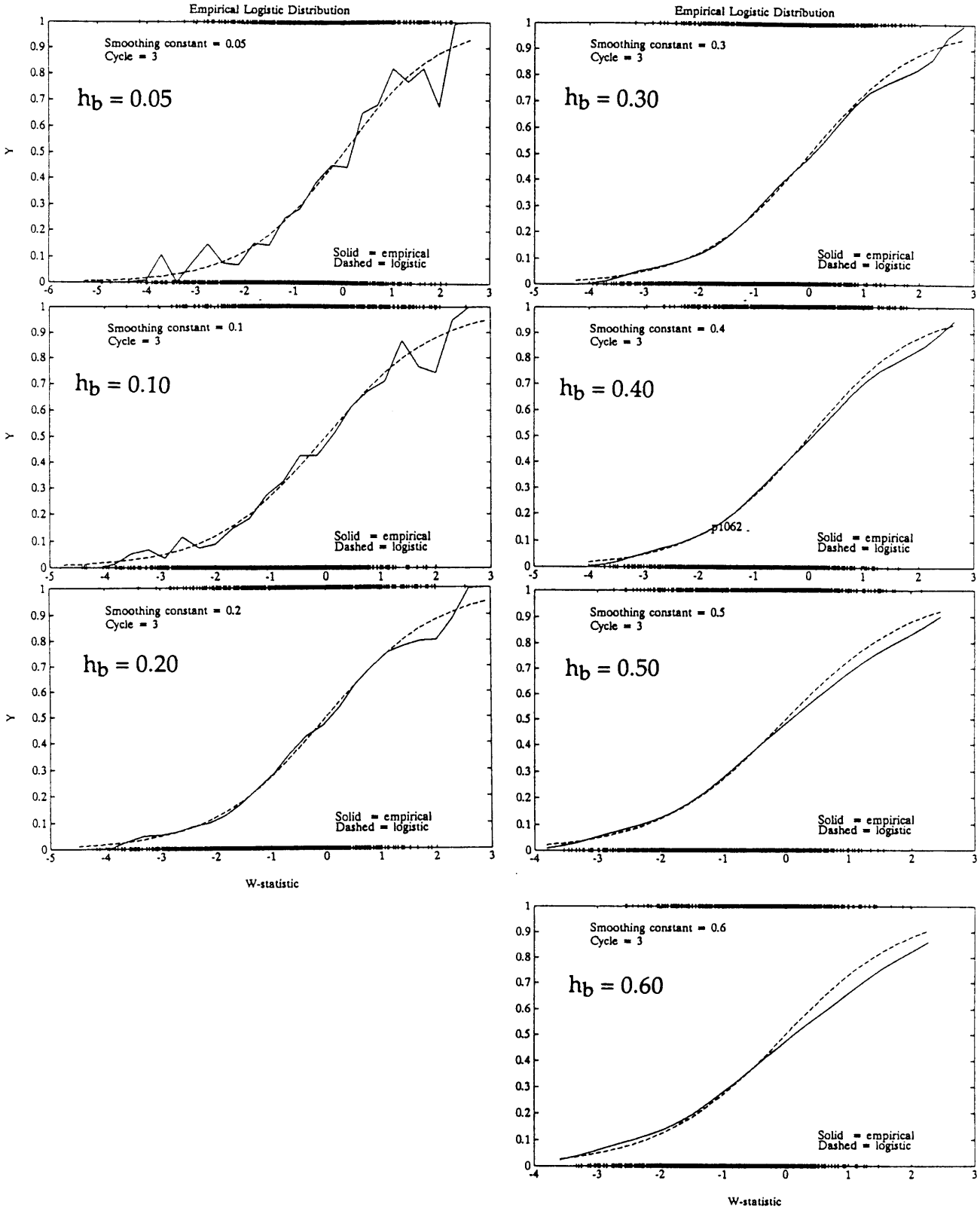


Figure 12: Empirical probability plots using different values of h_b

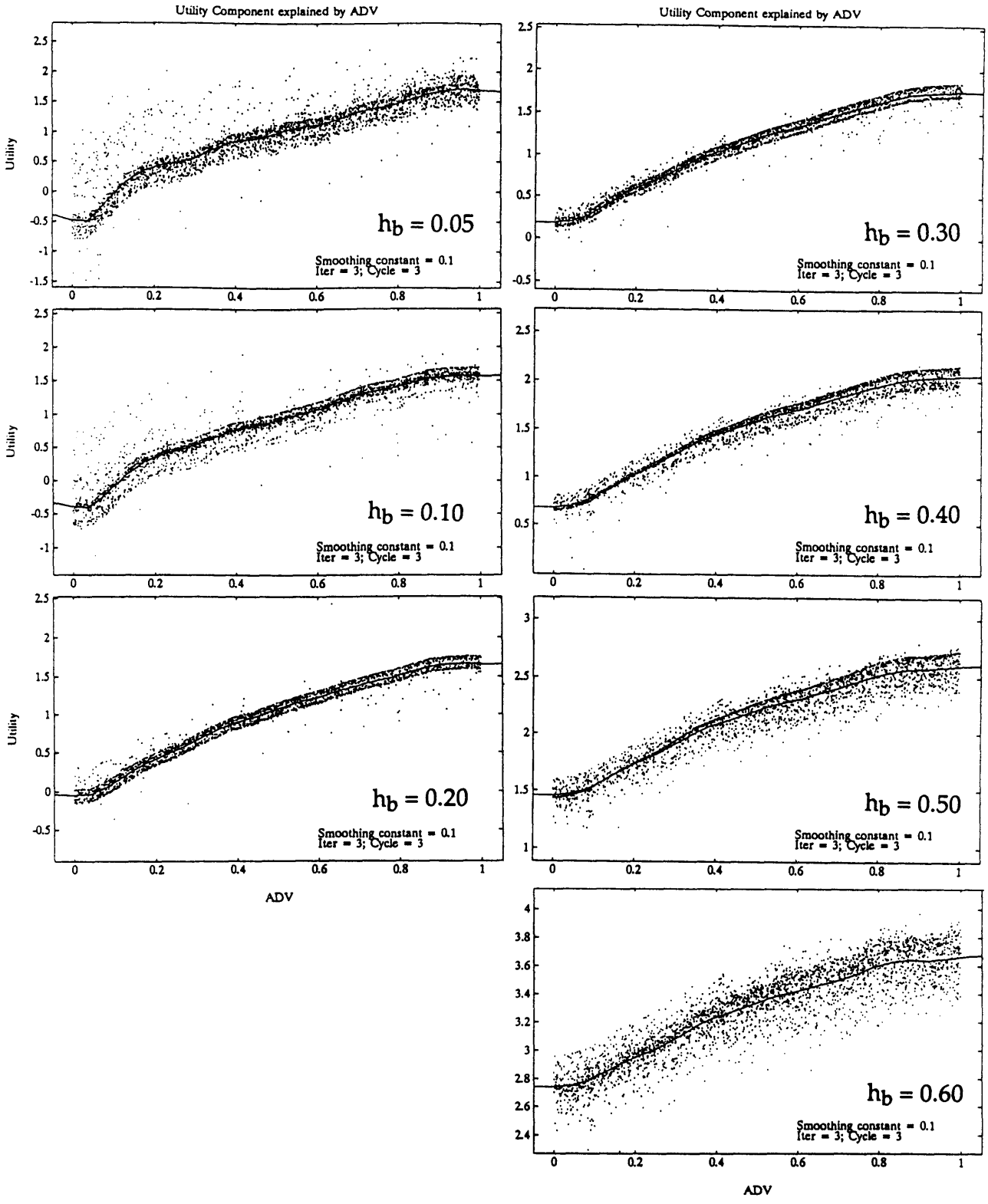


Figure 13: Advertising utility transformations using different values of h_b in the binary kernel regression

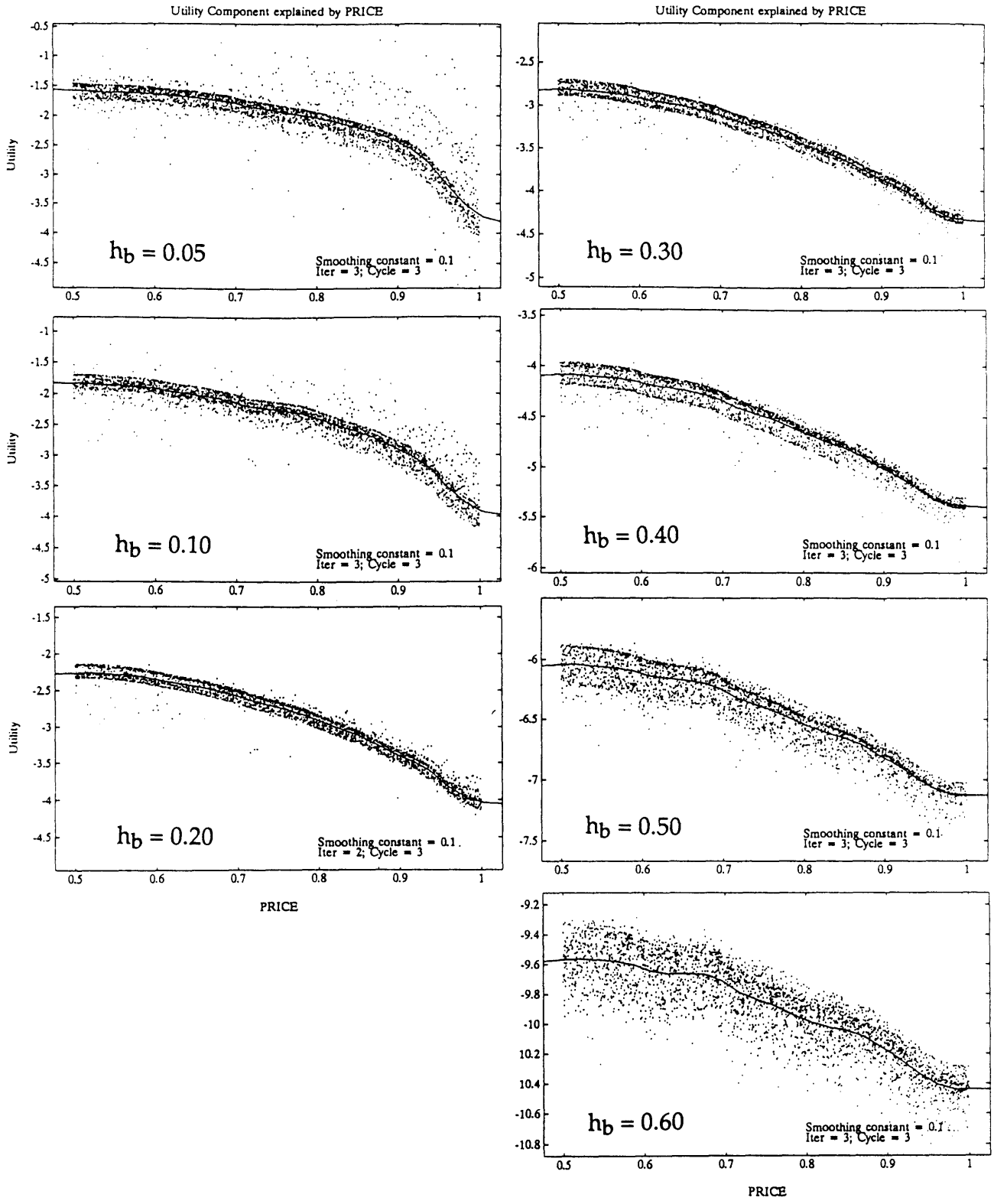


Figure 14: Price utility transformations using different values of h_b in the binary kernel regression

Another heuristic way for checking the oversmoothing in the probability plot is to examine the final MNL coefficients for the utility transformations. An empirical finding is that these coefficients increase systematically from near 1.0 for h_b=0.05 to as much as 2.0 for h_b=0.60. This is due to the result of the smoothing constant affecting the trade off between bias and variance. In general, bias caused by oversmoothing is something very hard to detect because the true underlying value of what is being estimated is unknown. However, in the case of the transformation coefficients, their true values are known to be 1.0. Thus, oversmoothing can be detected if they are much larger than 1.0. For our simulated dataset, recovery of the original models seems quite good for h_b=0.30 which has the coefficients of about 1.14.

4. APPLICATIONS TO SCANNER DATABASES

The simulation study in Section 3 illustrates that URM is quite successful in recovering the underlying additive utility structure and could potentially become a valuable tool by extending MNL models. In this section, the method is applied to real databases to [1] better understand market response, and [2] exploit possible advantages of URM over linear-in-parameters MNL.

4.1 Red Drink Single Source Database

This database contains 513 panel purchase records as well as advertising exposure data monitored by TV meters from 140 households in Grand Junction, Colorado. The products are so called red drinks, which includes cranberry or any blend of cranberry such as cranberry apple and cranberry grape. The five highest share brandsizes (share figures in parentheses) --- brand O cranberry cocktail 32oz (23.2%), 48oz (24.4%), 64oz (20.9%), cranapple 48oz (14.6%), and private label cranberry cocktail 48oz (17.0%) --- are extracted. They cover about 40% of the category purchases, and shares of other brandsizes are much smaller.[5]

[5] The decision to ignore smaller volume brandsizes is primarily for the computational convenience, and its consequence needs further investigation. Here, we simply argue that product switching to and from the selected five major brandsizes with other drink category is more frequent than with smaller share brandsizes because of diverse nature of the red drink category. Hence, small volume brandsizes are of less concern in our brand choice model.

The television commercial broadcasts (three types) for brand O are primarily brand image oriented and do not strongly differentiate among flavors. Thus, we assume the same advertising exposure level for all brand O brandsizes regardless of their flavor and size. There was no private label TV ad. An advertising variable for each panelist was constructed as a sum of all previous exposures encountered before a particular purchase occasion in question, adjusted for memory recall by daily decay. The decay rate was set to 0.95, which is equivalent to reaching approximately 20% level after 4 weeks based on studies by Lodish (1971), Clarke (1976), Craig, Sternthal, & Leavitt (1976). From anticipation that the advertising might prompt increased purchase quantity, advertising variables are introduced as interactions with size dummies to capture their size dependent effects.

In addition to the size specific advertising variables, loyalty, price per ounce, feature, and display are included. Although an initial MNL showed all advertising variables, Adv32 (t=-0.97), Adv48 (t=0.79), and Adv64 (t=-0.11) to be insignificant, after dropping Adv32, Adv48 was marginally significant (t=1.76) and Adv64 had t=0.86. This implies that the advertising may be causing an upgrading of the size purchased from 32oz to larger ones as well as brand switching within 48oz from private label to brand O. As a first cut, we have included only Adv48 in URM model.

Figure 15 illustrates the nonparametric utility functions of price, loyalty, and Adv48 by URM after the convergence is achieved after three iterations. By looking at these plots, a manager might make following observations.

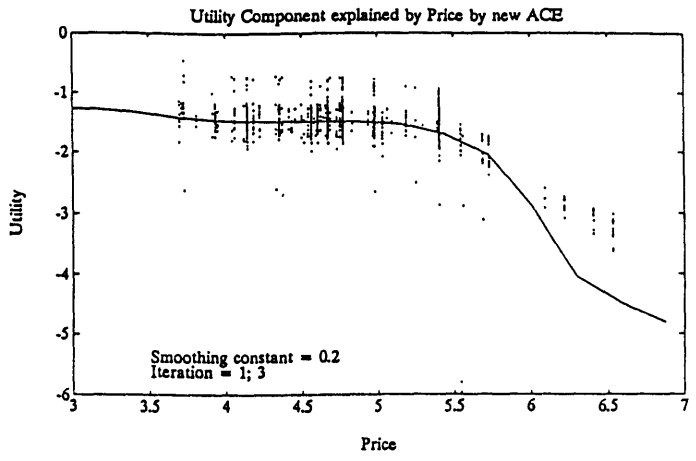
- There is a steep utility decrease for unit price above 5.7 cent/oz.[6]
- There is an evidence for the existence of three loyalty segments: brand non-users with loyalty less than 0.05, switchers between 0.05 and 0.9, and brand loyal users above 0.9.
- The advertising exposure exhibits saturation and possibly an S-shape.

Logit results for the linear model and URM are presented in Table 4. The fit is greatly improved as can be seen from the U-square jumping from 0.637 to 0.670. Note also that the advertising coefficient (df=1) has changed from insignificant (t=0.72) to significant (t=2.6) at α=1%.

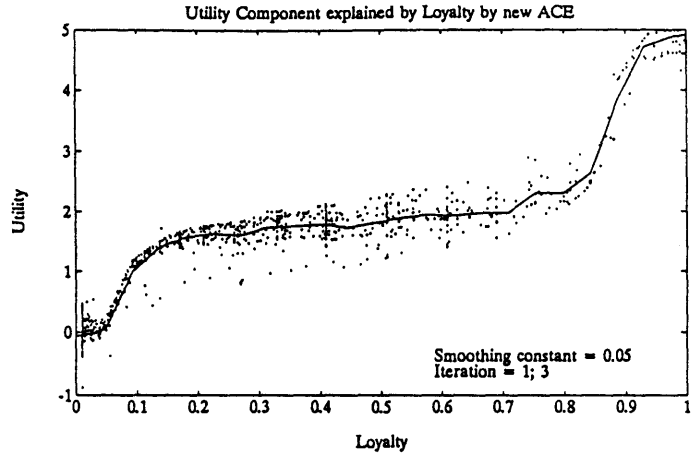
---

[6] Although the line does not seems to go through the points for price near 6 cents, there exist some points below the x-axis scale excluded from the graph.

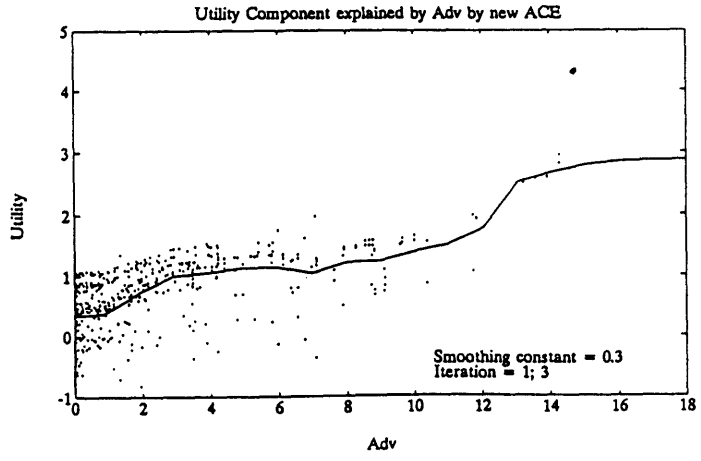
Utility residual plots



· Steep utility decrease for price above 5.7 cent/oz.



· Possibly three loyalty segments: non-users, switchers, and loyal users.



· Advertising response exhibits saturation and may be S-shaped.

Figure 15: Nonparametric utility transformations by URM in the Red Drink study

Table 4: Result of linear and URM logit in Red Drink Study

LINEAR MODEL

variable	coeff.	std.err	t-stat
ASC2	-0.282	0.236	-1.195
ASC3	-0.046	0.374	-0.122
ASC4	0.0061	0.275	0.022
ASC5	-0.500	0.277	-1.808
LOYALTY	4.256	0.216	19.727
FEATURE	1.744	0.500	3.487
DISPLAY	1.448	0.392	3.689
PRICE	-0.372	0.346	-1.074
ADV48	0.039	0.055	0.718

U-squared = 0.6370
Mean Absolute Deviation = 0.0486
Mean Absolute Second Derivative = 0.4417

URM NONPARAMETRIC TRANSFORMATION

variable	coeff.	std.err	t-stat
ASC2	-0.542	0.207	-2.620
ASC3	-0.448	0.292	-1.537
ASC4	0.056	0.250	0.225
ASC5	-0.878	0.276	-3.183
Φ(LOYALTY)	1.147	0.0750	15.299
FEATURE	1.701	0.501	3.394
DISPLAY	1.495	0.376	3.980
Φ(PRICE)	0.586*	0.502	1.167*
Φ(ADV48)	1.131	0.444	2.551

U-squared = 0.6696
Mean Absolute Deviation = 0.0152
Mean Absolute Second Derivative = 0.4757

*** the coefficient is not significant**

Finally, Figure 16 shows the empirical probability plot constructed from the utility residuals before and after URM is applied. The discrepancy from the theoretical value decreases markedly. Numerical values, the mean absolute deviation, appear in Table 4.

4.2 Aseptic Drink Database

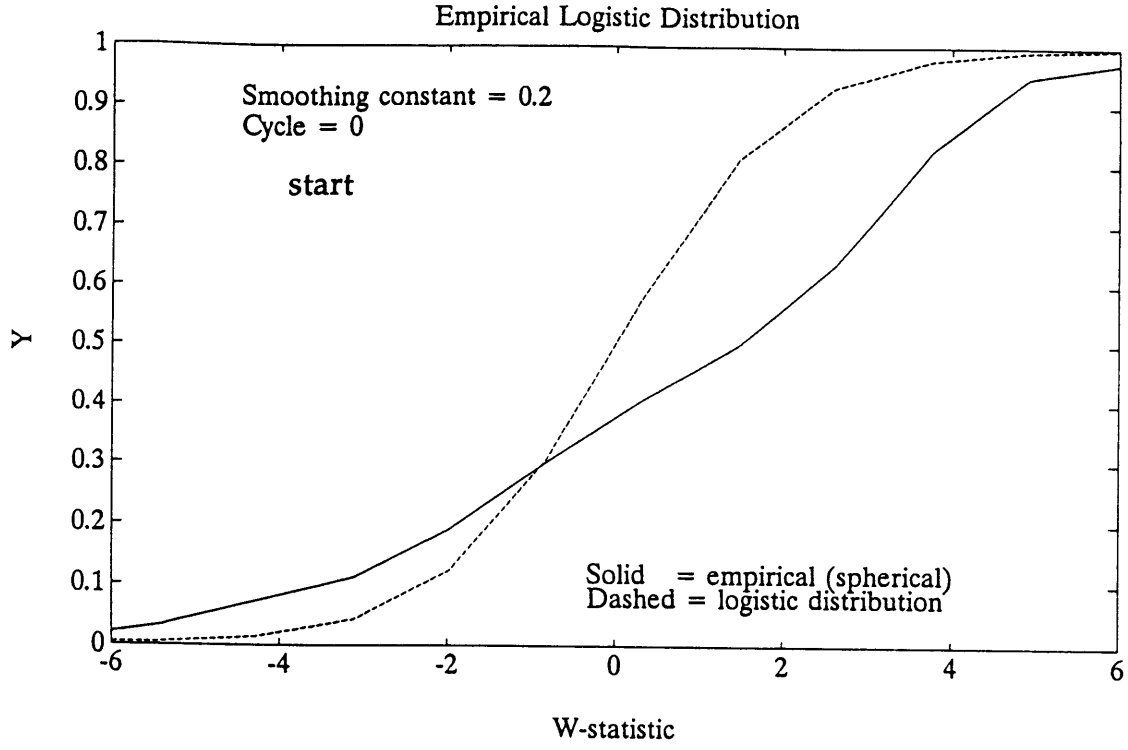
We now apply URM to scanner data that is a subset of the Aseptic Drink Database used in the study of the nonparametric density estimation (NDE) in Part I. Apart from adding more experience with URM, there are two purposes in this analysis. First, the model developed by URM is applied to a holdout sample in order to examine its predictive fit for model validation. Second, its fit and prediction are directly compared to the standard multinomial logit with linear utility and doubly exponential distributional assumption and to an NDE model which relaxes the distributional assumption as well as the entire concept of utility maximization.

The database contains purchase records for three major brands (Hi-C, KoolAid, Ssips) of three-pack aseptic drinks from 33 panelists during weeks of 12-29-86 through 2-6-89 (111 weeks). Category purchase occasions (observations) were separated according to those made before and after 5-9-88, which resulted in 988 observations in the calibration and 572 in the holdout sample.

For the comparison of MNL, URM, and NDE, marketing variables incorporated are limited to brand loyalty as defined in Guadagni & Little (1983), feature (0/1 binary), and display (0/1 binary) due to a large memory requirement of NDE. For details of NDE, refer to Part I.

Figure 17 shows the nonparametric function of loyalty and price obtained by URM.[7] The loyalty transformation indicates a quite nonlinear curve, while the price variable exhibits the discontinuous nature briefly mentioned in Section 3. Figure 18 is a time series share tracking of each brand by aggregating purchase occasions within each 4-week period for the actual data and prediction by MNL and URM. Table 5 shows these goodness-of-fit numerically, where the upper figures are for the calibration and lower ones are for the holdout.

[7] The price variable is included only in this figure. Any other comparison with NDE is done by excluding the price variable from the URM and MNL for fair comparison.



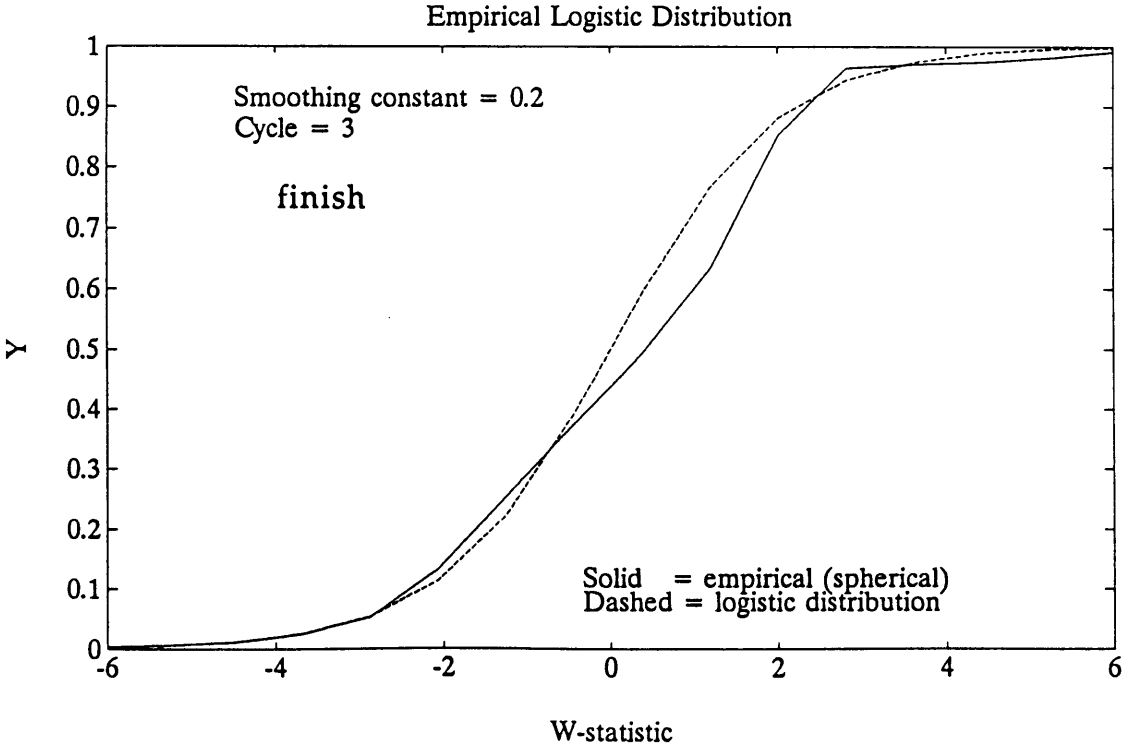
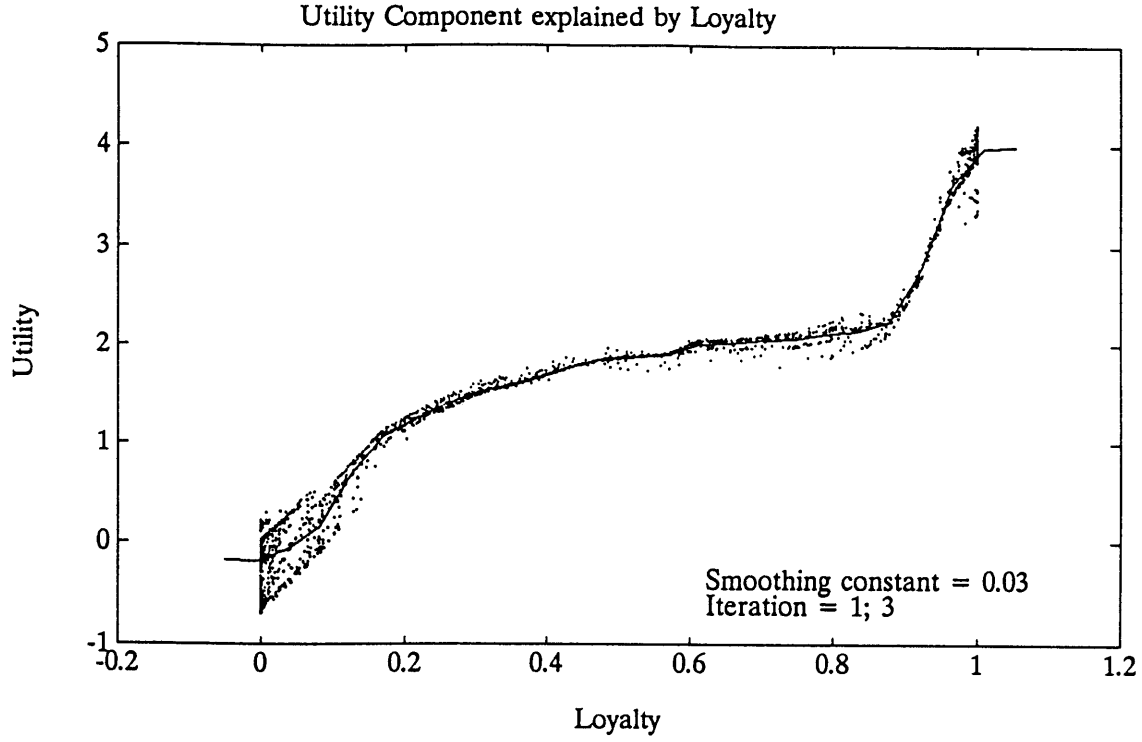


Figure 16: Empirical and theoretical probability plots in the Red Drink study



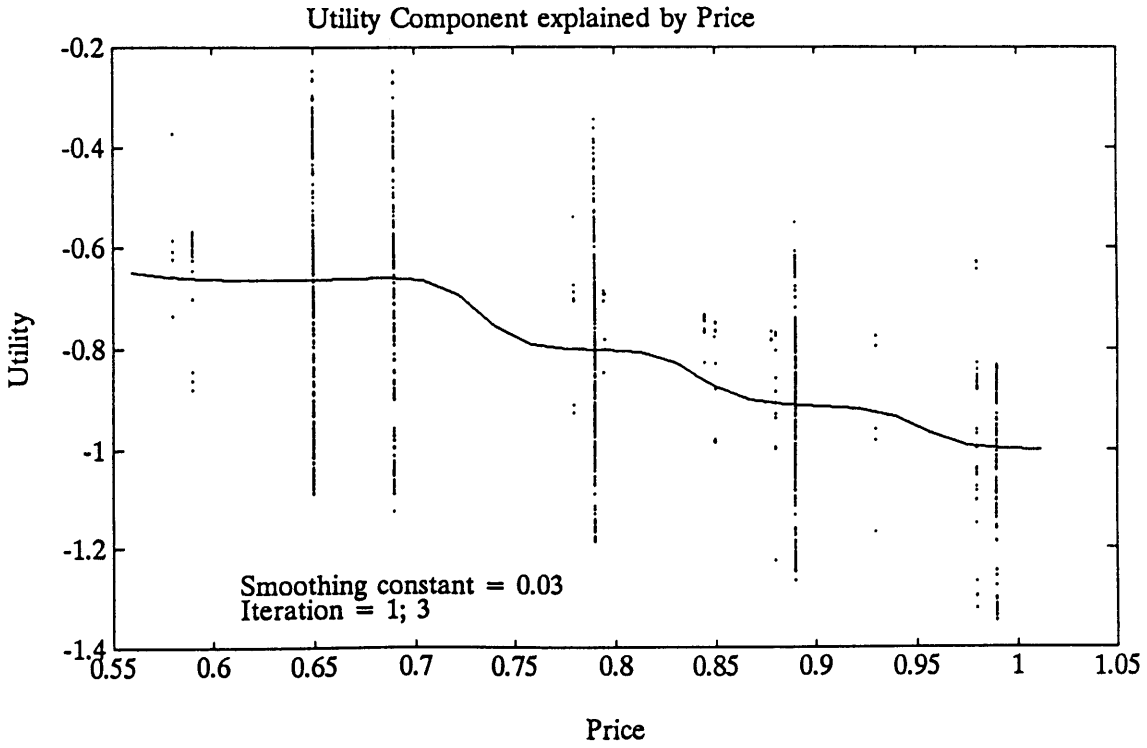
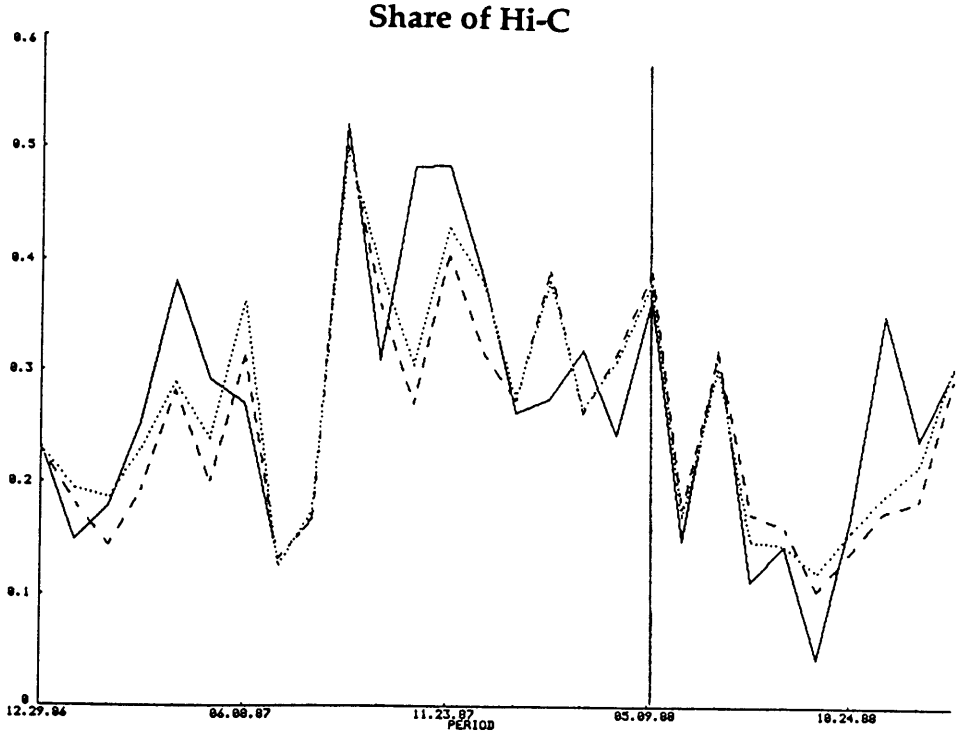


Figure 17: Additive nonparametric utility transformations by URM in Aseptic Drink study



solid = observed dash = URM model dotted = linear MNL model

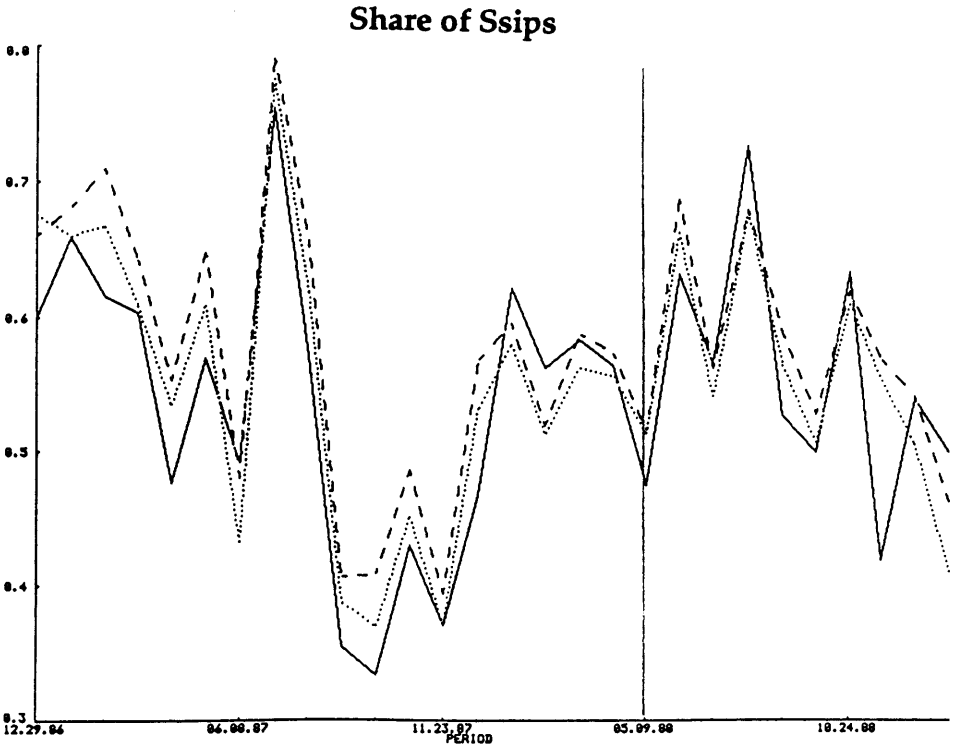


Figure 18: Time series share tracking of Hi-C and Ssips

Table 5: Goodness-of-fit in calibration and holdout sample

criterion		MNL		URM		NDE
Ave. loglikelihood	*calibration*	-0.4847	<	-0.4619	<	-0.4075
	holdout	-0.6050	<	-0.5678	>	-0.6850
R^2	*calibration*	0.8113	<	0.8322	<	0.8788
	holdout	0.8110	<	0.8290	>	0.7965

As might be expected in the calibration sample, the fit is improved in going from MNL to URM to NDE. This is because the effective degrees of freedom are increasing and fewer assumptions are made by the model. In the holdout, this is not necessarily true and in fact, we find that NDE seems to exhibit an overfitting phenomenon. However, this is not observed in URM. As sample size is increased, the improvement in goodness-of-fit is expected to increase more for models involving nonparametric techniques than for MNL. In fact, NDE dominated MNL in holdout with a larger Aseptic Drink database of 1,988 calibration observations in Part I.

Thus in this database, URM does very well. It fits better than MNL in the calibration sample and predicts better in the holdout. Relative to the computationally intensive and data thirsty NDE approach, URM does not fit as well in the calibration sample (which NDE seems to overfit) but outperforms NDE in more important task of predicting in the holdout sample.

5. SUMMARY

At this stage of its development, URM represents a potentially valuable exploratory tool for discrete choice models. The simulation study demonstrates that the method recovers the underlying additive utility structure quite well and is robust under various distributions for the stochastic utility. Pointwise confidence intervals for the reconstructed nonparametric utility functions are shown to indicate reliability of the estimators. Also, a notion of the degrees of freedom is derived to facilitate the selection of the smoothing constants and statistical

inferences. The two scanner data applications suggest the usefulness and advantages of URM in studying market response and shed light on modeling practice --- a trade-off between strength of assumptions and the model performance in cross-validation.

From the current study, it seems fair to conclude that URM can offer much improvement over the linear MNL when the utility function is additive and nonlinear. It has better prediction, as expected, since a URM model generalizes the linear specification of a MNL model by nonparametric functions. The utility transformations are intuitively appealing because they permit visual inspection of nonlinear response functions, and one has always an option of recurring to parsimonious parametric modeling after inferring an appropriate functional form. URM offers the promise of becoming a tool to assist managers make more effective marketing decisions.

There are two possible directions for extending URM. In substantive marketing applications, URM could be useful for observation and confirmation of various marketing response phenomena predicted by behavioral and perceptual theories. Advertising responses (Kanetkar, Weinberg, and Weiss 1989, Tellis 1988) and reference price (Winer 1986, Kalwani et al. 1989, Lattin and Bucklin 1989, Gurumurthy and Little 1989) are some of the interesting areas where URM could be valuable. In some cases, it might be useful to set up alternative specific nonparametric utility functions analogous to alternative specific parameters in MNL to capture difference in elasticities and responsiveness among alternatives.

Mathematically, there exists many questions to be answered for URM to be a standard statistical tool, not just an exploratory device. For example, how does the order of variables to be cycled affect the result? My computational experience indicates that the resulting utility functions are rather insensitive to the order after a few cycles. But are the estimated nonparametric utility functions consistent when the additivity assumption holds? Understanding of the relevant literature in ACE (Breiman and Friedman 1985) and GAM (Hastie and Tibshirani 1986, 1987) and seeking their connection with URM may help give us some answers. It would be desirable to determine whether the estimated nonlinear utility functions are consistent in the statistical sense. These and other questions await further research.

APPENDIX A

Derivation of Theoretical Probability Plot, E(y | w), by Utility Maximization

In MNL, the stochastic utility associated with alternative j is expressed as

$$u_j = v_j + \varepsilon$$

where ε is an i.i.d. random variable of Gumbel (doubly exponential) distribution with $E(\varepsilon)=0$ and $Var(\varepsilon)=\pi^2/6$. If necessary, the scale and the level of v_j can be normalized to meet this condition. A distributional form of ε is

$$F(\varepsilon) = \exp\left(-e^{-(\varepsilon+\gamma)}\right) \qquad \text{where } \gamma \text{ is Euler constant } (\approx 0.577)$$

Then, the probability of choosing alternative j from choice set J in MNL is

$$P(j) = \text{Prob}\,(u_j > u_i\ ;\ "\ i \in J,\ i \neq j) \qquad \text{(A1)}$$

$$= \text{Prob}\,(u_j > MAX_{i\in J,\, i\neq j}\, u_i) \qquad \text{(A2)}$$

$$= \text{Prob}\left\{v_j + \varepsilon_j > MAX_{i\in J,\, i\neq j}\,(v_i + \varepsilon_i)\right\} \qquad \text{(A3)}$$

$$= \text{Prob}\left\{v_j + \varepsilon_j > \ln \sum_{i\neq j} e^{v_i} + \varepsilon^*\right\} \qquad \text{(A4)}$$

$$= \text{Prob}\left\{v_j - \ln \sum_{i\neq j} e^{v_i} > \varepsilon^* - \varepsilon_j\right\} \qquad \text{(A5)}$$

$$= \text{Prob}\left\{v_j - \ln \sum_{i\neq j} e^{v_i} > \xi\right\} \qquad \text{where } \xi = \varepsilon^* - \varepsilon_j \quad \text{(A6)}$$

Equation (A4) holds because if u_m and u_n are independent random variables of Gumbel distribution with $E(u_m)=v_m$, $E(u_n)=v_n$, and $Var(u_m)=Var(u_n)=\pi^2/6$, then $\max(u_m, u_n)$ is also Gumbel distributed with mean $\log(e^{u_m} + e^{u_n})$ and variance $\pi^2/6$. (Ben-Akiva and Lerman, 1985) Thus, ε^* is also Gumbel distributed with mean 0 and variance $\pi^2/6$. Furthermore, it can be easily shown (Ben-Akiva and Lerman, 1985) that ξ, which is a difference of two independent Gumbel distributed random variables of the same mean and variance $\pi^2/6$, is logistically distributed as

$$F(\xi) = \frac{1}{1+e^{-\xi}} \qquad \text{(A7)}$$

$F(\xi)$ has a smooth S-shape and is symmetric about a point $F(\xi=0) = 0.5$.

Now, because $P(j)=E(y_j \mid w_j)$, where y_j is 1 if alternative j is chosen, 0 otherwise, (A6) can be expressed as

$$E(y_j \mid w_j) = \text{prob}(\xi < w_j), \qquad \text{where } w_j = v_j - \ln \sum_{i \neq j} e^{v_i}$$

In other words, a regression line of y_j on a single explanatory variable w_j should depict a cumulative distribution function of the logistic distribution ξ shown in (A7) if the error term satisfies the distributional assumption of MNL.

APPENDIX B

Derivation of Approximate Degrees of Freedom in Kernel Regression

Kernel regression to compute $E(y \mid x)$ is given as

(B1)
$$E(y \mid x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)}$$

where $K(\cdot)$ is a kernel function, h is a smoothing constant, and n is the number of observations. It can be interpreted as a local weighted averaging of y_i by rewriting (B1) as

$$E(y \mid x) = \sum_i \lambda_i y_i$$

(B2)
$$\text{where} \quad \lambda_i = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_j K\left(\frac{x - x_j}{h}\right)} \quad \text{and} \quad \sum_i \lambda_i = 1$$

It is a linear estimator which can be expressed as $\hat{\mathbf{f}} = \mathbf{S}\,\mathbf{y}$, where $\hat{\mathbf{f}}$ and $\mathbf{y}$ are $n \times 1$ vector of fitted and observed response variable, and S is a $n \times n$ matrix. In terms of the kernel regression weights above, each row of S corresponds to a $1 \times n$ row vector λ whose elements consist of λ_i shown in (B2) evaluated at each observation. Hence, the i-th diagonal element of S, S_{ii} is

(B3) $$S_{ii} = \frac{K(0)}{\sum_j K\left(\frac{x_i - x_j}{h}\right)}$$

As discussed in the main text, we define the degrees of freedom as trace of S, i.e., ΣS_{ii} . If n is sufficiently large and x_i's are reasonably spread out, there is a simple analytic formula for the trace as a function of the smoothing constant h in the case of a single explanatory variable and a Gaussian kernel function.

Hastie and Tibshirani (1990) note that the smoothing constant is the major determinant of the degrees of freedom while the predictor configuration has little effect. Hence, a good approximation can be obtained by considering n observations of (x_i, y_i), where x_i are uniformly spaced between 0 and 1 so that the interval between adjacent x_i's is $\Delta x=1/n$. Gaussian kernel function is

$$K(z) = \frac{1}{\sqrt{2\pi h_s^2}} \exp\left(-\frac{z^2}{2h_s^2}\right)$$

For large n, the numerator and denominator of equation (B3) becomes,

$$K(0) = \frac{1}{\sqrt{2\pi h_s^2}} \Delta x$$

and $$\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h_s}\right) \approx \frac{1}{\sqrt{2\pi h_s^2}} \int_0^1 \exp\left(-\frac{(x_i - x)^2}{2h_s^2}\right) dx$$

The latter is simply an area between 0 and 1 under the normal distribution of variance h_s^2 centered at x_i. By substituting them into (B3),

$$S_{ii} = \frac{1}{\sqrt{2\pi h_s^2}} \frac{1}{n} \frac{1}{\text{Area}(x_i; h_s)}$$

For fixed h_s, the area takes the minimum value $A(h_s)$ at $x_i = 0$ and 1, and the maximum value $B(h_s)$ at $x_i = 0.5$, where

$$A(h_s) = \Phi\left(\frac{1}{h_s}\right) - \frac{1}{2} \qquad \text{and} \qquad B(h_s) = 2\,\Phi\left(\frac{0.5}{h_s}\right) - 1$$

Now, the trace of S is

$$\begin{aligned} tr(S) &= \sum_i S_{ii} \\ &= \frac{1}{\sqrt{2\pi h_s^2}} \frac{1}{n} \sum_i \frac{1}{Area(x_i; h_s)} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{h_s} \alpha(h_s) \end{aligned} \tag{B4}$$

where $\alpha(h_s)$ is a mean of the reciprocals of the area, which occurs somewhere between $1/B(h_s)$ and $1/A(h_s)$. It is numerically evaluated and showed in the following Table B.

Table B: Numerically Evaluated $\alpha(h_s)$

h_s	$\alpha(h_s)$
0.01	1.01273
0.05	1.05954
0.07	1.08295
0.10	1.11806
0.20	1.23516
0.30	1.35562
0.40	1.49439
0.50	1.65973
0.60	1.84738
0.70	2.05090
0.80	2.26539
0.90	2.48757
1.00	2.71530

To illustrate the formula, let us compute the value of h_s^* which corresponds to the degrees of freedom equal to 1. One would expect h_s^* to be of the order of 1 such that the kernel regression takes almost global average in the domain of x. Using equation (B4) and Table B, $h_s^* \approx 0.9$. The fact that this value of h_s^* leads to near global averaging can be seen by plotting a normal distribution curve of standard deviation h_s^* centered at 0.5. The weights at the boundaries (x = 0 and 1) retain 86% of that in the center.

To apply this formula to the smoothing constant h defined on an arbitrary range of the domain of x, h must be scaled to the standardized h_s by the ratio of the standard deviation of observed x_i's, σ_x, to that of the uniform distribution above, namely, $\sqrt{1/12}$. In other words, h_s must be

substituted by $h_s = (\sqrt{1/12})h/\sigma_x$. Therefore, the general approximation formula for the degrees of freedom is

(B5) $$tr(S) = \frac{1}{\sqrt{2\pi}} \frac{1}{h} \frac{\sigma_x}{\sqrt{1/12}} \alpha\left(\frac{h\sqrt{1/12}}{\sigma_x}\right)$$

where $\alpha(\cdot)$ is shown in Table B.

In this paper, all h is normalized with respect to $\sigma_x=1$. Thus, it is simply,

(B6) $$tr(S) \approx \frac{1.389}{h} \alpha(0.2887\, h)$$

and plotted in Figure B.

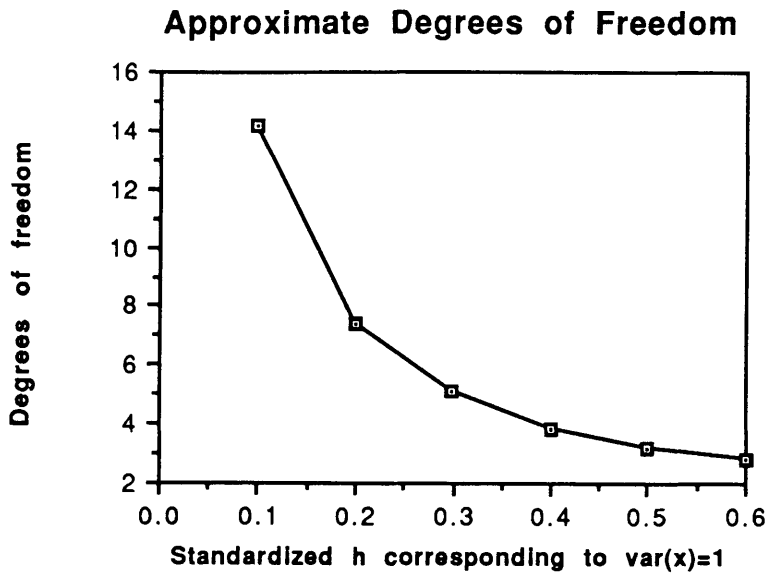


Figure B: Approximate Degrees of Freedom v.s. Smoothing Constant

Comparison with the actual trace of S indicates that equation (B6) is indeed a good approximation. In our simulation study, trace of S for the advertising and price transformation regression for h=0.1 are 14.33 and 14.43 respectively, while the degrees of freedom using (B6) is 14.3.

APPENDIX C

Extension of the Generalized Additive Models

C.1 Brief Review of Generalized Linear Models (GLM)

GLM (Nelder and Wedderburn 1972) relate a random variable y to a vector of explanatory variables **x**, but relax some assumptions associated with standard regression. They generalize many likelihood based regression models to a common framework. The models consist of three components;

[1] Random component

The response y is independent identically distributed with exponential family density,

$$\text{(C1)} \qquad f_Y(y;\theta,\phi) = \exp\left\{ \frac{y\,\theta - b(\theta)}{a(\phi)} + c(y,\phi) \right\}$$

where θ is called the natural parameter, and ϕ is the dispersion parameter. Expectation and variance of y is shown to be $E(y)\equiv\mu=\partial b(\theta)/\partial\theta$ and $Var(y)\equiv v=\partial^2 b(\theta)/\partial\theta^2 \cdot a(\phi)$. Because our concern is how the mean of y is related to the explanatory variables, without loss of generality the dispersion is assumed to be independent of **x** and equal to $a(\phi)=1$ by letting θ absorb the scaling.[8]

[2] Systematic component

A linear predictor, η, is related to a vector of explanatory variables **x** as $\eta=\mathbf{x}'\beta$.

[3] The link

η is related to the mean of y by a link function, $g(\cdot)$, as $\eta = g(\mu)$.

Different specifications of the exponential distribution and link function create various regression type models. For example, using a normally distributed random component and an identity link function, $\eta = \mu$, produces a standard regression model. Bernoulli distribution can be obtained by substituting $a(\phi)=1$, $b(\theta)=\log(1+e^{\theta})$, and $c(y,\phi)=0$ in (C1). In this case, the link

[8] This practice is fairly standard as in McCullagh and Nelder (1989) and Hastie and Tibshirani (1990).

function must be chosen such that the domain of $g(\cdot)$, i.e. the probability $\mu \in [0,1]$, is mapped to the entire real axis of $\eta = x'\beta$. If the inverse logistic function, so called logit function, $\eta = \log(\mu/1-\mu)$, is chosen, the model becomes binary logit while the inverse cumulative normal, $\eta = \Phi^{-1}(\mu)$, leads to a binary probit model. Similarly, multinomial distribution is expressed as a generalized linear model, and the use of inverse logistic and cumulative normal link function result in multinomial logit and probit respectively.[9]

The most commonly used link is called canonical link, $\eta = \theta$, and in this case there exists a sufficient statistic for β. (McCullagh and Nelder 1989) Since its derivation is not documented in their book, it is shown in Appendix D.

Once, each of the three components is specified, MLE of β can be found by the Fisher scoring procedure, a variant of the Newton-Raphson algorithm. It is shown in McCullagh and Nelder

---

[9] Explicit form of the multinomial distribution in GLM is illustrated below since it is not treated by McCullagh and Nelder (1989).

Consider the case of J alternatives Let $\mathbf{y}$ be a J×1 column vector of an outcome indicator (i.e. 1 for an element corresponding to the outcome and 0 otherwise). Now, the parameter of GLM, θ, is also a J×1 vector. Then, the joint probability mass function of a vector $\mathbf{y}$ is,

(F1) $f_y(\mathbf{y}) = \exp\{ \theta'\mathbf{y} - b(\theta) \}$ from (C1)

Define $b(\theta) = \log (\sum_j e^{\theta_j})$

Then, (F1) becomes the familiar

$$f_y(\mathbf{y}) = \frac{e^{\theta'\mathbf{y}}}{\sum_j e^{\theta_j}}$$

with $$\mu_i(\theta) = E(y_i) = \frac{\partial b(\theta)}{\partial \theta_i} = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}} .$$

The canonical link, $\theta = \eta = x'\beta$ leads to MNL as

$$\mu_i(\beta) = f_{y_i}(y_i) = \frac{e^{x_i'\beta}}{\sum_j e^{x_j'\beta}}$$

Multinomial probit is obtained by the inverse cumulative normal link, Φ, as

$$\eta_i = \Phi^{-1}(\mu_i) \quad \text{or} \quad \mu_i = \Phi(x_i'\beta)$$

(1989) that the estimation process is equivalent to adjusted dependent variable regression (in a sense that they produce the identical estimation sequence), where the adjusted dependent variable z derived from the current estimate of $\hat{\mu}$ and $\hat{\eta}$ is regressed on explanatory variables x with weights w defined as

$$z = \hat{\eta} + (y - \hat{\mu}) \left(\frac{d\eta}{d\mu}\right)_{\mu=\hat{\mu}} \qquad \text{(C2)}$$

$$w^{-1} = \text{var}(y) \left(\frac{d\eta}{d\mu}\right)^2_{\mu=\hat{\mu}} . \qquad \text{(C3)}$$

The algorithm can be regarded as iteratively estimating a new β by the weighted least squares regression of the linear predictor η on explanatory variables x, in which η is updated by the Taylor series expansion around μ based on the current estimate of β. Iteration terminates when the increase in loglikelihood is sufficiently small.

C.2 Brief Review of Generalized Additive Models (GAM)

GAM (Hastie and Tibshirani 1986) extends the systematic component of GLM, η, to a sum of one dimensional nonparametric functions, $f_p(x_p)$ in each explanatory variable x_p as $\eta = \Sigma f_p(x_p)$ rather than the linear-in-parameters form. Definitions for the other two components of GLM, the random component and the link function, are unchanged. The following operational modifications on GLM are introduced to compute the more general form of η. First, the estimation of each coefficient with the least square criterion is substituted by the estimation of functions using one dimensional nonparametric regression (more generally referred to as a smoother by Hastie and Tibshirani). Second, one step estimation of a vector of coefficients, β, by the weighted least squares regression is now a backfitting procedure (Friedman and Stuetzel 1981), where the nonparametric functions f_p's are estimated sequentially, one variable at a time, using the previous estimates of functions for other variables. The backfitting algorithm is summarized as follows.

Initialization by linear model: $f_p(x_p) = \beta_p x_p \quad \forall\, p = 1,..,P$

Cycle over explanatory variables: $p = 1,..,P, 1,..,P, 1,..$

$$f_p(x_p) = E\{ y - \Sigma_{q \neq p} f_q(x_q) \mid x_p \}$$

Until changes in the functions, $f_p(x_p)$*, are sufficiently small.*

Third, the iteratively reweighted least squares (IRLS) algorithm for η is replaced by so called the local scoring algorithm in which Fisher's scoring is updated using a local score estimate. The adjusted dependent variable z for GAM is derived by the Newton-Raphson method as

(C4) $$z = E\left[\hat{\eta} + (y - \hat{\mu}) \left(\frac{d\eta}{d\mu}\right)_{\mu=\hat{\mu}} \middle| x\right]$$

where $\hat{\mu}$ and $\hat{\eta}$ are computed based on the current estimate.

Hastie and Tibshirani (1986) demonstrate that the local scoring algorithm achieves the maximum expected loglikelihood. Let us denote the additive predictor in the systematic component as $\eta(x)$ and a loglikelihood function for response y as $L(\eta,y)$. Then, the estimate, $\eta = \Sigma f_p(x_p)$, maximizes $E[L(\eta,y)]$, instead of the sample loglikelihood, $\Sigma_i L(\eta_i,y_i)$, as is the case in MLE. This criterion has an intuitive appeal in that the estimate maximizes the loglikelihood for all future observations. Maximizing the sample loglikelihood in GAM produces non-smooth estimated functions. For instance, in the binomial case, the estimated functions will be such that $\eta(x_i)=+\infty$ for $y_i=1$ and $\eta(x_i)=-\infty$ for $y_i=0$. A sketch of the derivation for the local scoring algorithm is shown in Appendix E.

Technical discussion for convergence of the algorithm and consistency (existence) and non-degeneracy (uniqueness) of the solution can be found in Buja, Hastie, and Tibshirani (1989). The following informally summarizes their findings.

For most common nonparametric regression methods satisfying some general technical conditions,

[1] there exists at least one solution (i.e. f_p $\forall p$) to which the algorithm converges.

[2] if the concurvity space, analogous to multicollinearity in functional space and null space in linear systems, is empty, the solution is unique. Otherwise, the algorithm converges to one of the solutions depending on the starting functions.

One basic example of concurvity in a function is represented by a base level or intercept term. Thus, if the solution, say $f_1(x_1)$ and $f_2(x_2)$, is not constrained to be centered to mean 0, it leads to degeneracy because $f_1(x_1) + k$ and $f_2(x_2) - k$ is also a solution. In practice, as long as covariates are not highly correlated in the sense of concurvity, convergence of the algorithm and consistency of the solution are generally achieved.

C.3 GAM Applied to Logit Models

Binary logit, or more commonly known as logistic regression, is one of GAM studied by Hastie and Tibshirani (1986, 1987) with a Bernoulli random component and a canonical link, $\eta(x)=g(\mu)=\log(\mu/1-\mu)$. It can be written in a more familiar form as

(C5) $$\mu(x) = E(y \mid x) = \frac{1}{1 + e^{-\eta(x)}}$$

(C6) $$\text{where } \eta(x) = \Sigma f_p(x_p)$$

The adjusted dependent variable of equation (C4) without the conditional expectation becomes

(C7) $$z = \eta(x) + \frac{y - \mu}{\mu (1 - \mu)}$$

where $\eta(x)$ and μ are evaluated based on the current estimate. The additive functions f_p's are obtained by nonparametric regression of z on x with weights, $w=\mu(1-\mu)$, by the backfitting algorithm. Equation (C7) can be interpreted as computing the first order Taylor series approximation of the linear predictor, $\eta = g(\mu)$, from its current estimate. A direct approach of using the observed response, y, to obtain $\eta=g(y)$ does not work here because of its binary nature. It is note worthy that if $\eta(x)$ is linearly specified as $x'\beta$ in equation (C7), z is what is referred to as partial residuals by Landwehr, Pregibon, and Shoemaker (1984).

The following is the local scoring algorithm for logistic regression.

Initial estimate by linear model, $\eta(x)=x'\beta$

Repeat

Compute the current estimate of μ *from* η [by (C5)]

Compute the adjusted dependent variable z *and the weights* w. [by (C7)]

Obtain $f_p(x_p)$'s *by nonparametric regression of* z *on* x *with weights* w.

[by the backfitting algorithm]

Until loglikelihood converges.

The goal of this paper is to obtain a multinomial logit model with an additive utility function as

(C8) $$P_j=\frac{e^{v_j}}{\sum_k e^{v_k}} \qquad \text{where} \qquad v_j=\sum_p \phi_p(x_{jp})$$

In the case of only two alternatives, the above logistic regression for GAM can be adapted with slight modifications. Denoting the probability of choosing alternative 1 as $\mu(x)$, (C8) can be reduced to the form of (C5) with

(C9) $$\eta = v_1 - v_2 \quad \Rightarrow \quad \Sigma_q f_q(x_q) = \Sigma_p \{ \phi_p(x_{p1}) - \phi_p(x_{p2}) \}.$$

There are two possible approaches in formulating equation (C9). One is to assume that only a difference of explanatory variables influences the probability but not their absolute levels, which is a standard linear-in-parameter logit assumption. In equation (C9), this implies that $x_q = x_{p1} - x_{p2}$ and q=p. The other is to define alternative specific utility functions so that

$$\begin{cases} f_q(x_q) = \phi_p(x_{p1}) & q = 1, \ldots, P \\ f_q(x_q) = -\phi_p(x_{p2}) & q = P+1, \ldots, 2P \end{cases}$$

This is analogous to alternative specific coefficients in the standard logit.

For more than two alternatives, however, such simple solutions do not apply since the additive predictor is no longer additive in each covariate. That is,

(C10) $$\eta(x) = v_1(x) - \log \sum_{j \neq 1} e^{v_j(x)}$$

This would violate the fundamental assumption of GAM in the binomial formulation. Nevertheless, one can go back to the spirit of the basic approach of GAM and re-derive the appropriate formulation from a penalized conditional likelihood function. Because even a sketch of the derivation is rather complex involving much conceptual development and new terminology, here an algorithm for MNL is simply presented below. It is derived by re-interpreting the matched case-control model of Hastie and Tibshirani (1990, section 8.2) and making the following conversions. Cases are changed to chosen alternatives, and a pool of controls are replaced by non-chosen alternatives. The lower case k which ranges from 1 to K becomes an index for purchases, while the lower case r which takes a value from 1 to R_k+1 becomes and index for alternatives at the k-th purchase. Thus, K is the number of purchases and R_k+1 implies the number of alternatives in the choice set for the k-th purchase. We occasionally emphasize the distinction from the standard GAM by referring to it as the extended GAM, if necessary.

Initial estimate by linear model, $\eta(x_i)=x_i'\beta \quad \forall\, i$

Repeat

Compute the current estimate of μ_i *from* η *as*

$$\mu_i = \frac{e^{\eta(x_i)}}{\sum_j e^{\eta(x_j)}}$$

Compute the adjusted dependent variable z_i *and the weights* w_i*, where*

$$z_i = \eta(x_i) + \frac{y_i - \mu_i}{\mu_i (1 - \mu_i)}$$

$$w_i = \mu_i (1 - \mu_i)$$

Obtain $\phi_p(x_p)$*'s by nonparametric regression of* z_i *on* x_i *with weights* w_i.

[by the backfitting algorithm]

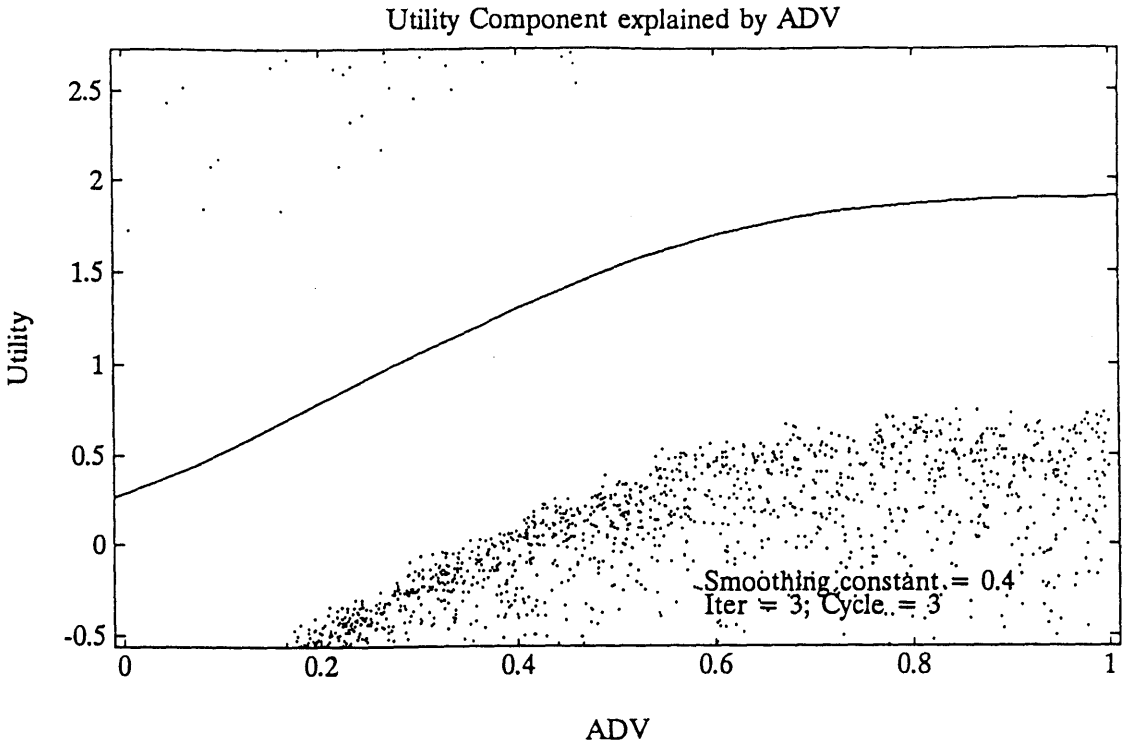
Until loglikelihood converges.

C.4 Connection with utility residual method (URM)

The objective of GAM is same as URM in that they both estimate the additive nonparametric utility functions in the MNL framework as shown in (C8). The difference lies on their execution in inferring an unobserved utility (or predictor η with the notation of GAM) from the discrete response. In GAM, a new utility η_n is obtained from the first order Taylor series approximation of the logit link function, logit(y), about μ, the best current guess to $E(y \mid x)$. On the other hand, URM constructs an empirical version of the inverse link function based on the current estimate. And the new utility is computed such that the discrepancy from the theoretical inverse link, a logistic function, is minimized.

C.5 Simulation Study

The same simulation study as in Section 3 is conducted to investigate the operational characteristics of the extended GAM described in Section C.3. Figure C1 shows the resulting nonparametric additive utility functions for advertising and price after convergence is achieved in three iterations. The smoothing constants are subjectively chosen to be h=0.4, which corresponds to 3.9 degrees of freedom by formula (8). The logarithmic relationship for advertising and the negative cubic for price are recovered quite well, however, the similar boundary effects as URM exist. Note the way residuals are distributed. There is a band of blank observations around the fitted curve, which is almost reverse case of the URM. Figure C1 is rescaled to show all residuals in Figure C2. The phenomenon is a consequence of the formulation of the "residual", $(y-\mu)/\mu(1-\mu)$, in the extended GAM described in the algorithm. Because y is an observed binary choice and μ is a predicted probability, those points above the fit in the figures corresponds to y=1 while those under are for y=0. Landweher, Pregibon, and Shoemaker (1984) obtain similar partial residual plots when $\eta(x)$ is a linear-in-parameters form, which coincides with the first iteration of the GAM. In Figure C3, the updated utility, $\eta(x)+(y-\mu)/\mu(1-\mu)$, is plotted against the utility $\eta(x)$ based on the current estimate for each iteration. Again, the similar blank band can be observed.



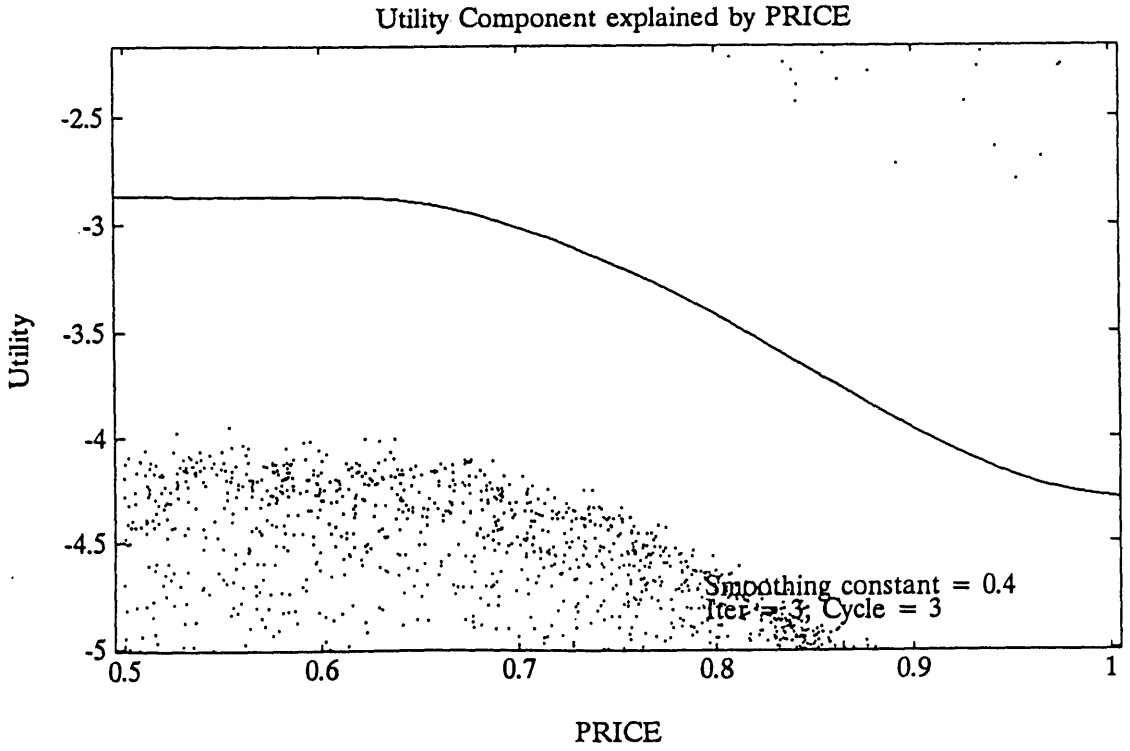
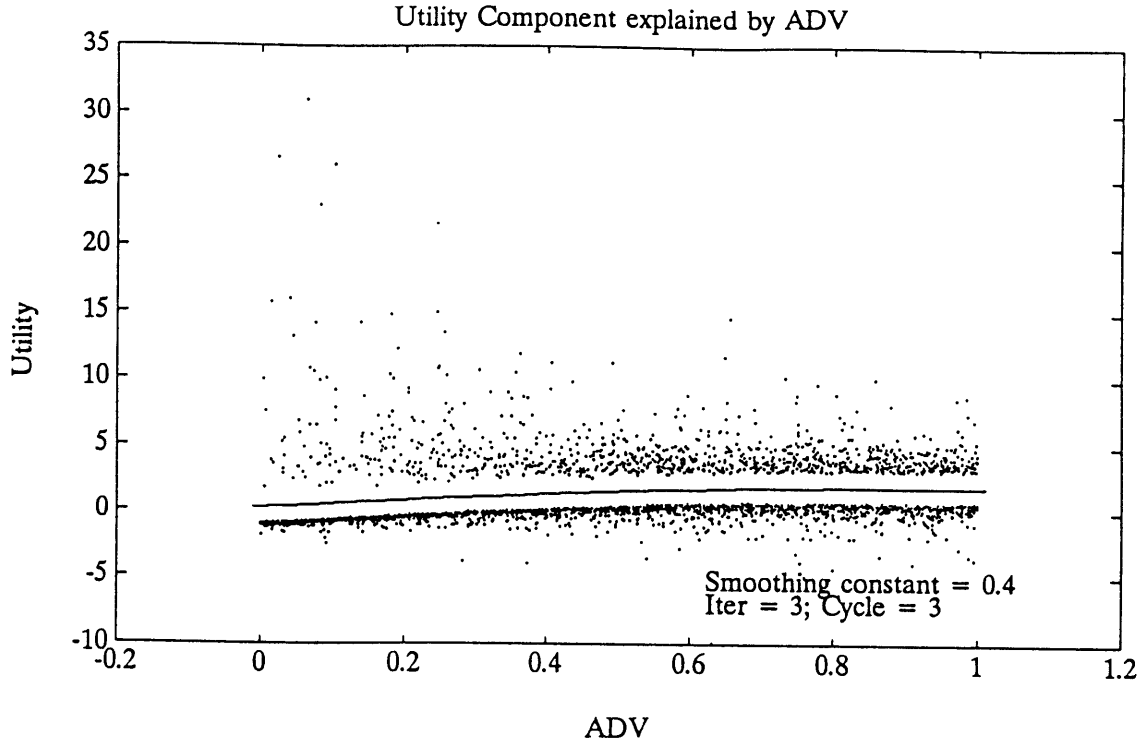


Figure C1: Additive nonparametric utility transformations by the GAM in the simulation study



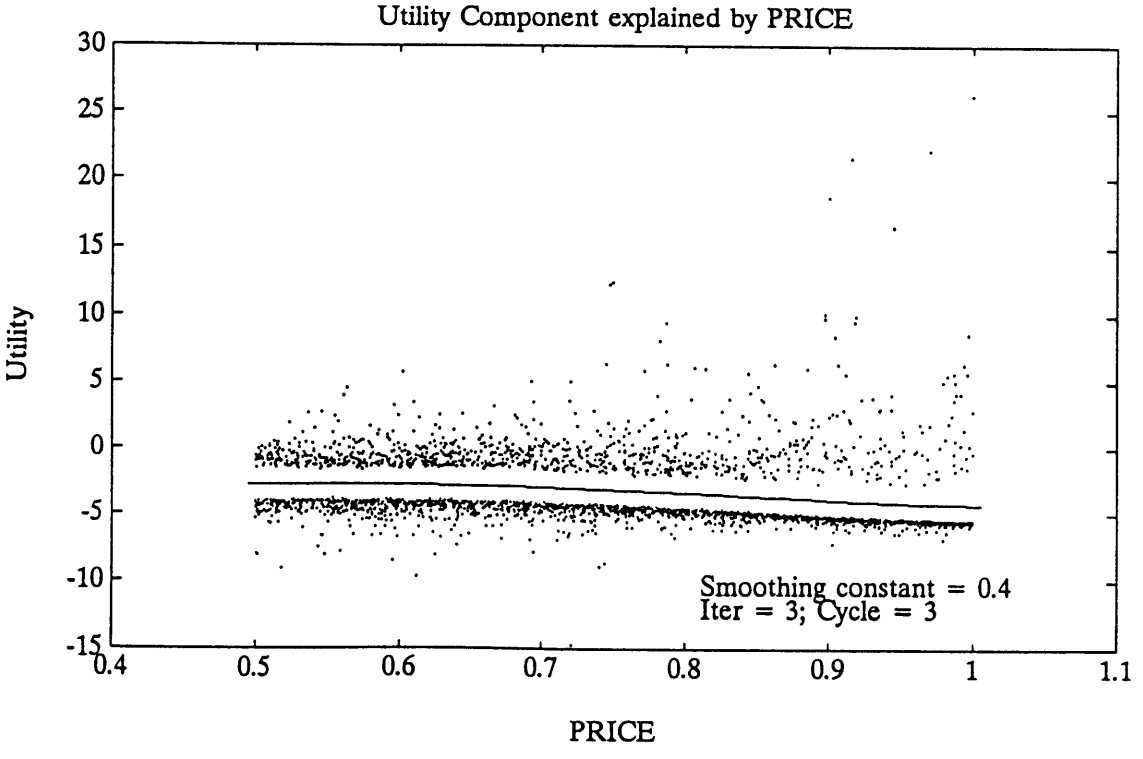


Figure C2: Rescaled additive nonparametric utility transformations by the GAM in the simulation study

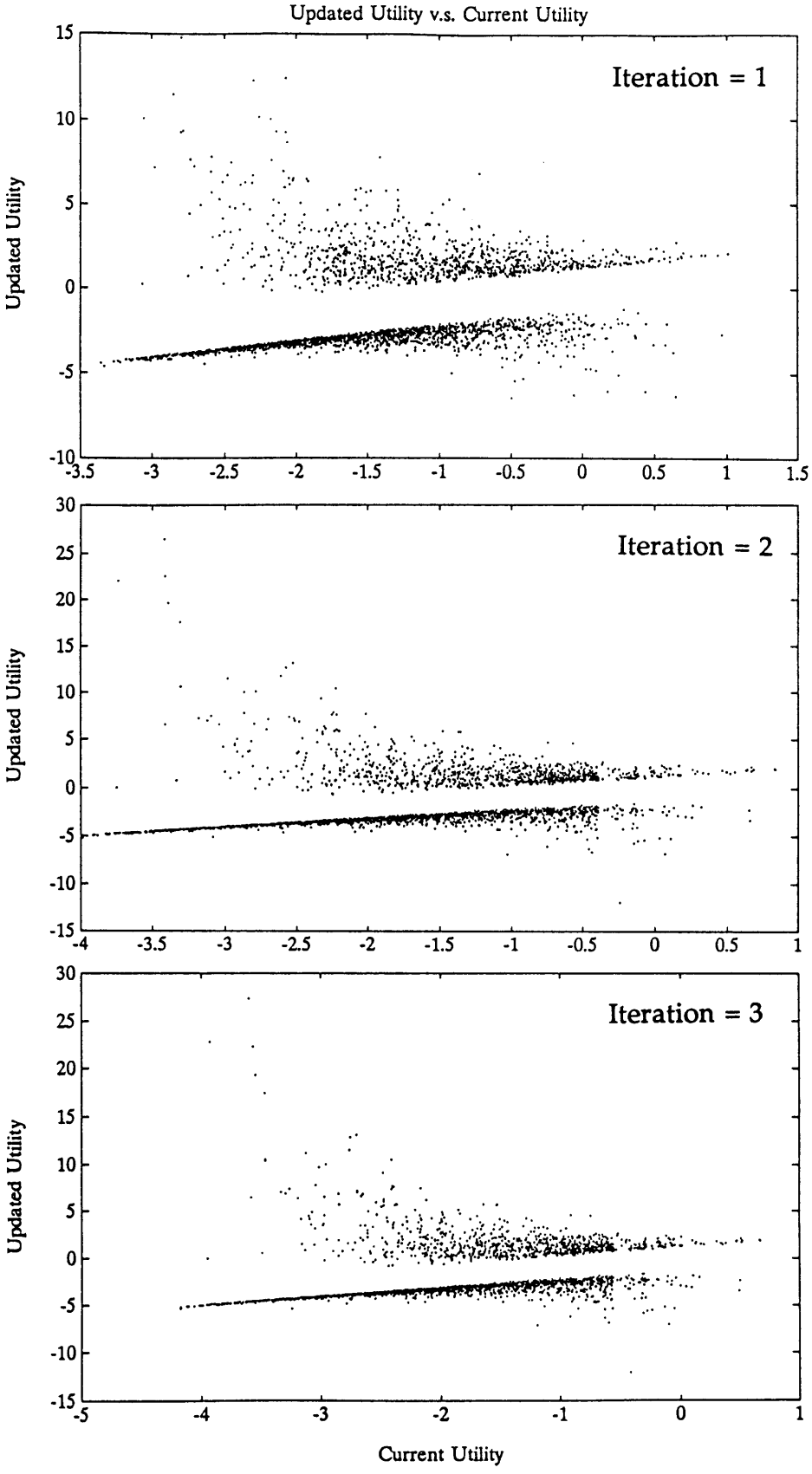


Figure C3: Updated utility v.s. current utility for each iteration

Table C1: Result of linear and the GAM logit in the simulation study

Linear model

variable	coeff.	std.err	t-stat
FEATURE	0.5811	0.1189	4.8850
DISPLAY	0.7681	0.2142	3.5863
ADV	1.8391	0.1585	11.6049
PRICE	-3.4936	0.3189	-10.9558
ASC2	0.2098	0.0932	2.2516
ASC3	0.2860	0.0918	3.1167

$\rho^2 = 0.14643$

The Extended GAM

variable	coeff.	std.err	t-stat
FEATURE	0.5833	0.1201	4.8560
DISPLAY	0.8295	0.2165	3.8321
ϕ(ADV)	1.1688	0.0966	12.1033
ϕ(PRICE)	1.1334	0.0980	11.5681
ASC2	0.2207	0.0942	2.3418
ASC3	0.3036	0.0926	3.2794

$\rho^2 = 0.16467$

The True Underlying Model

variable	coeff.	std.err	t-stat
FEATURE	0.5883	0.1199	4.9083
DISPLAY	0.8036	0.2157	3.7246
log(ADV+0.1)	1.0108	0.0848	11.9263
PRICE^3	-2.1616	0.1915	-11.2883
ASC2	0.2141	0.0938	2.2813
ASC3	0.2957	0.0923	3.2026

$\rho^2 = 0.16057$

Table C1 illustrates how well the GAM recovered the original functions by applying linear, the extended GAM, and the true model transformations of the advertising and price to MNL. The resulting ρ^2 actually exceeds that of the true model. Bootstrap estimate of 50 samples indicates 153.3% recovery in the loglikelihood value with a standard error of 4.78%. This overfitting is not surprising considering that the GAM is trying to maximize the expected loglikelihood by fitting even random noise into the structure of the nonparametric utility functions. Increasing the smoothing constants for both advertising and price from current value of 0.4 reduces ρ^2 and overfitting at the cost of biases in the resulting utility functions. Table C2 illustrates a sensitivity analysis of the smoothing constants in the range between 0.1 and 0.6. The included are loglikelihood value, ρ^2, and estimated scale coefficients for the transformations. Figures C4 and C5 show the resulting transformations for advertising and price respectively.

Table C2: Sensitivity Analysis for the Smoothing Constant

h	ρ^2	loglikelihood	coefficients (adv / price)
0.1	0.17813	-892.08	1.035 / 1.045
0.2	0.17019	-900.70	1.062 / 1.046
0.3	0.16641	-904.80	1.107 / 1.076
0.4 *	0.16467	-906.70	1.169 / 1.133
0.5	0.16372	-907.72	1.250 / 1.212
0.6	0.16294	-908.57	1.351 / 1.310

* The value chosen in the study

The computation time for the study varies from run to run depending on how many iterations are repeated. But in all cases, convergence is achieved within 5 iterations and actual time is between 3 and 5 minutes, which is about 40% less than that of URM.

Just as in URM, the extended GAM utilizes a logistic link function and is subject to the MNL distribution assumption. Therefore, we test robustness of the GAM with the same five distributions used in Section 3.2. Figure C6 and C7 illustrate the estimated advertising and

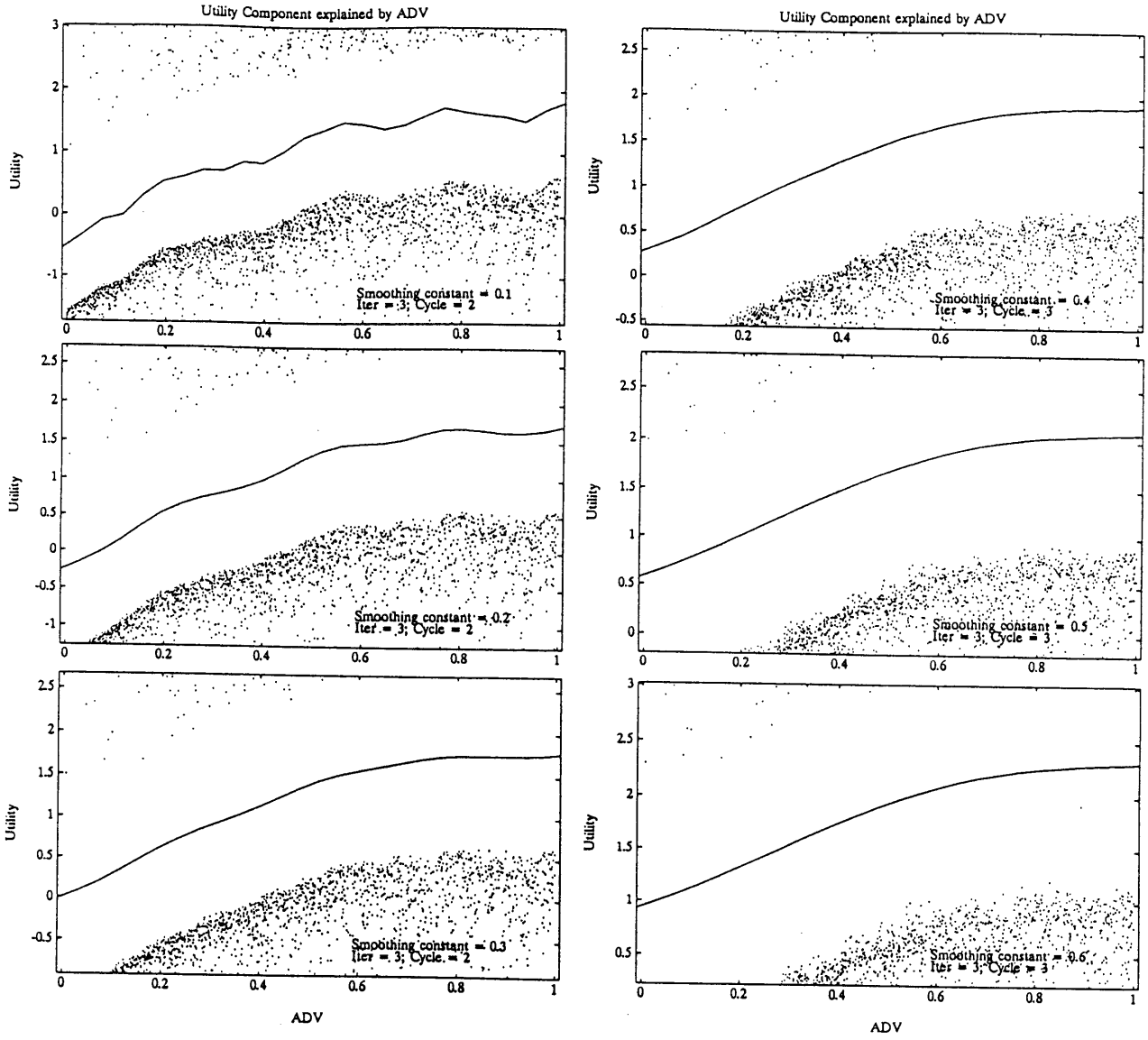


Figure C4: Advertising utility transformations using different values of the smoothing constant

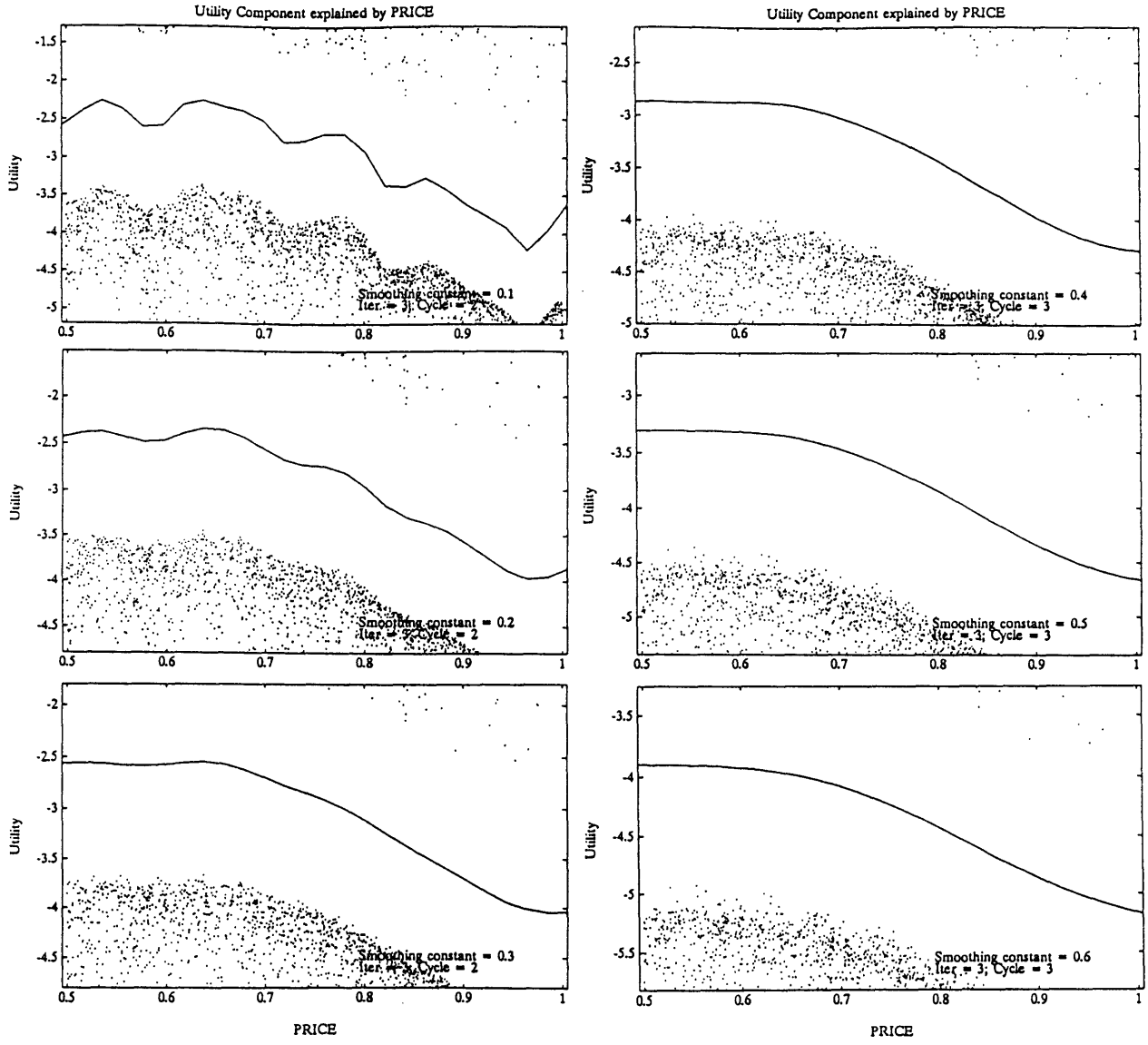


Figure C5: Price utility transformations using different values of the smoothing constant

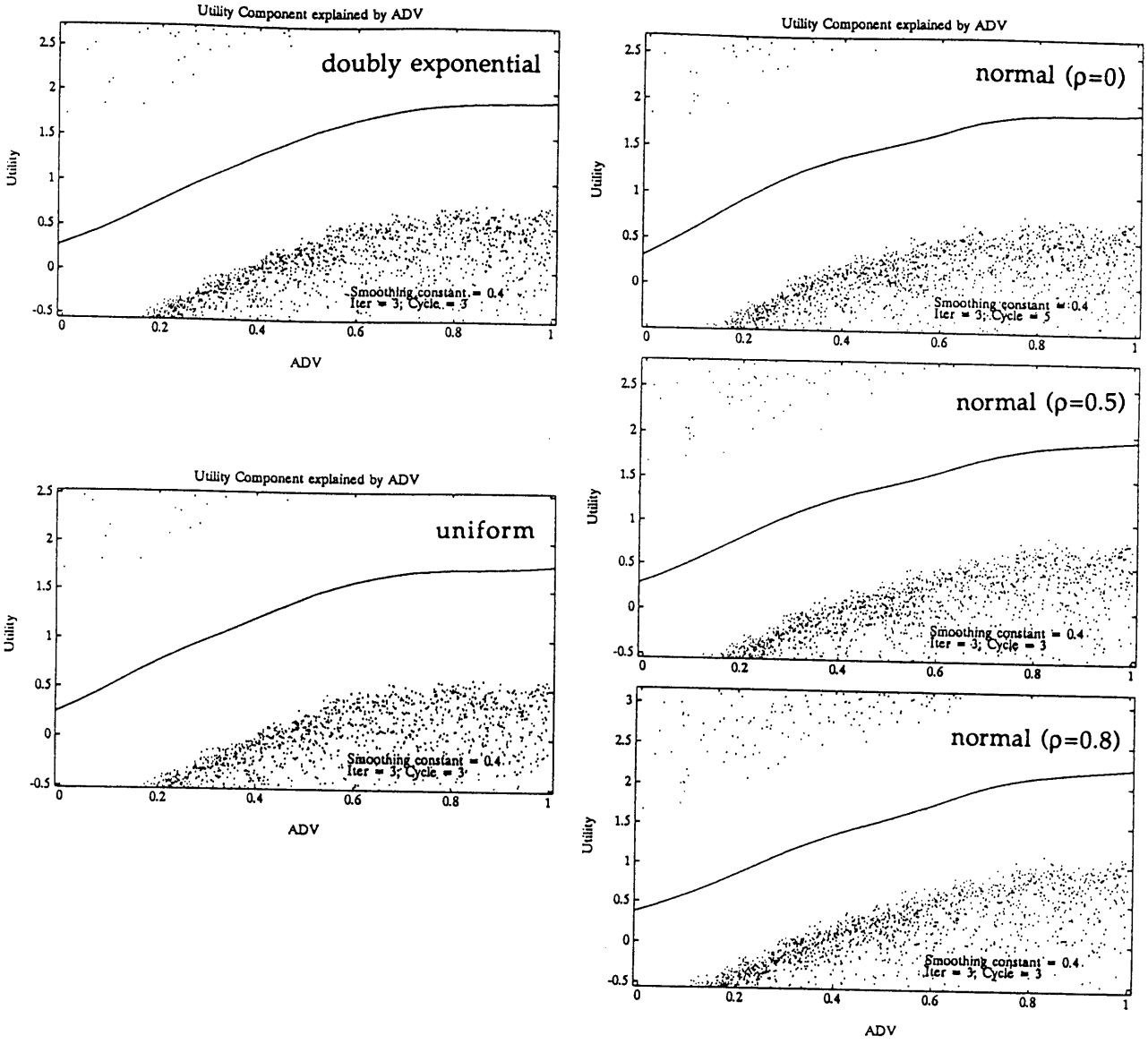


Figure C6: Advertising utility transformation with different error distributions

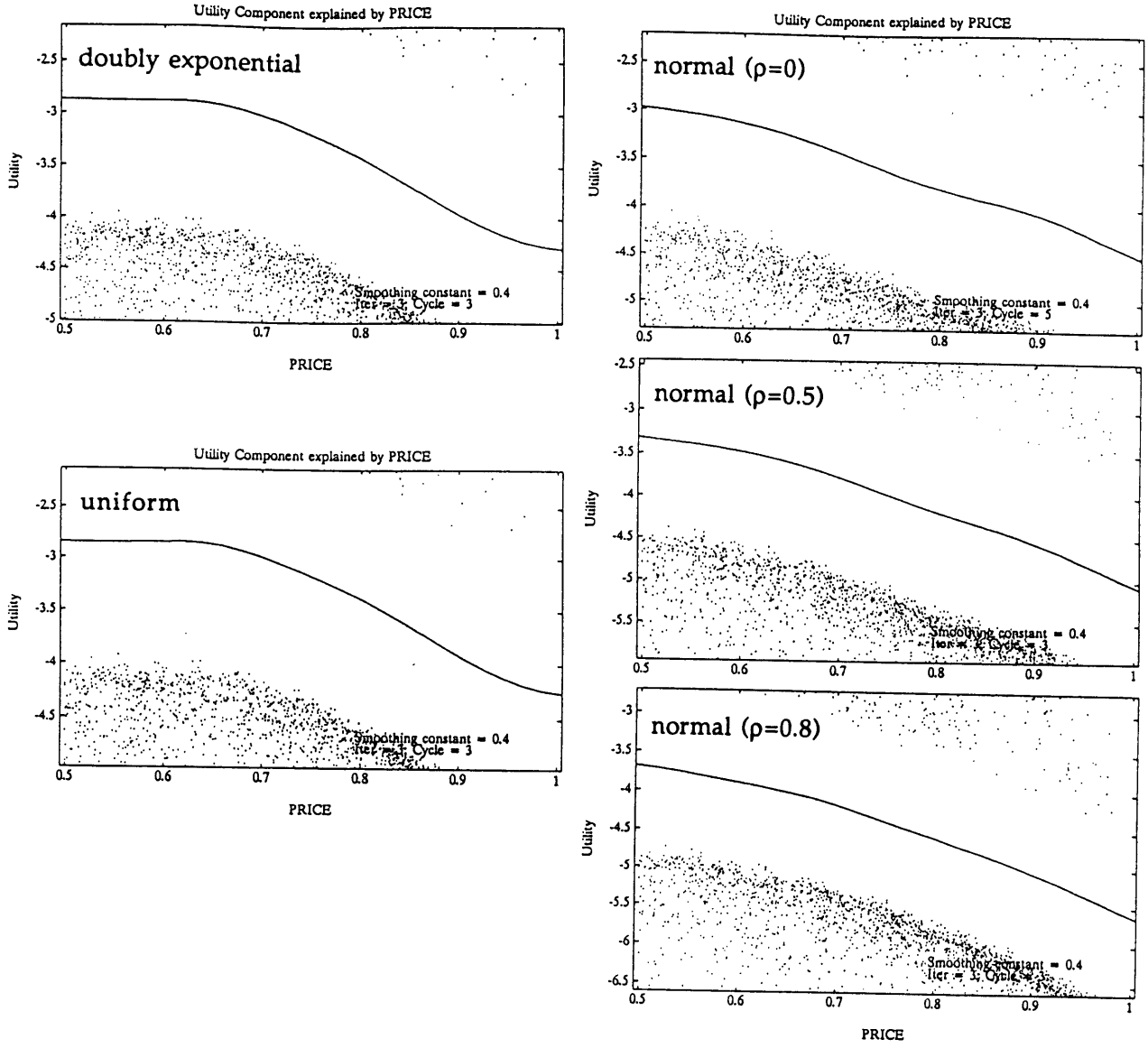


Figure C7: Price utility transformation with different error distributions

price utility functions respectively. Surprisingly, plots for the uniform distribution look more similar to the correctly specified doubly exponential ones that those for the independent normal, despite the fact that the normal distribution closely resembles the doubly exponential. Under a violation of error independence among alternatives, the GAM does not recover the underlying true utility structure as well as URM even for the correlation of 0.5. Because the result is based on the single sample, nothing affirmative can be concluded. However, URM seems to be more robust than the GAM. This is further supported by the operational failure of the GAM in some databases, which will be discussed in Section C.6.

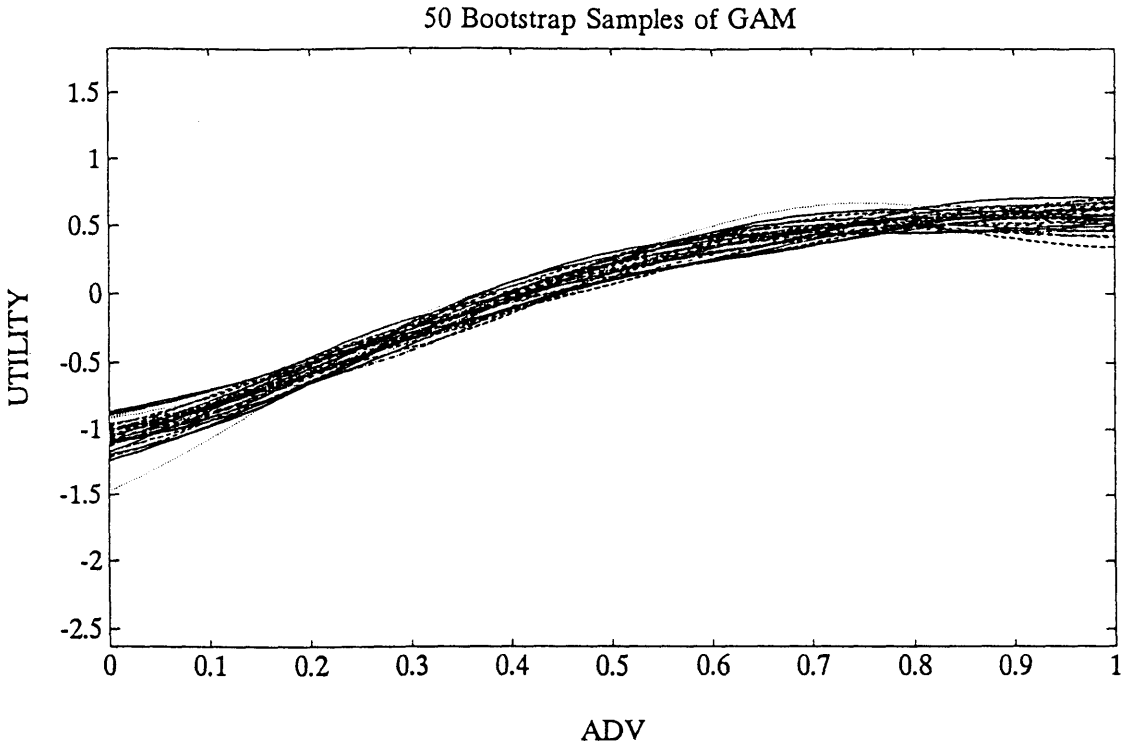
Table C3: Goodness-of-fit of Various Nested Models

model	deviance	degrees of freedom
asc2, asc3	2169.10	2
asc2, asc3, feature, display	2129.18	4
asc2, asc3, feature, display, price	2002.50	5
asc2, asc3, feature, display, adv	1985.02	5
asc2, asc3, feature, display, price, adv	1852.98	6
asc2, asc3, feature, display, ϕ_{price}	1831.78	7.9
asc2, asc3, feature, display, ϕ_{adv}	1833.72	7.9
asc2, asc3, feature, display, ϕ_{price}, ϕ_{adv}	1813.40	11.8

Goodness-of-fit for various nested models are shown in Table C3 with degrees of freedom using a formula (8) or (B5) derived in Appendix B. All three nonparametric utility models are significant compared with their linearly specified counterparts at a level of 1%. In addition, the nonparametric model with both ϕ(adv) and ϕ(price) is also significant in comparison to models with either ϕ(adv) or ϕ(price) alone at 1% level.

Figure C8 shows utility transformations of 50 bootstrap samples, from which 95% pointwise confidence bands are derived in Figure C9. They are slightly narrower than those by URM.

One of the reasons of choosing the kernel method for nonparametric regression is that it can be easily modified to accommodate multiple explanatory variables. In higher dimensional



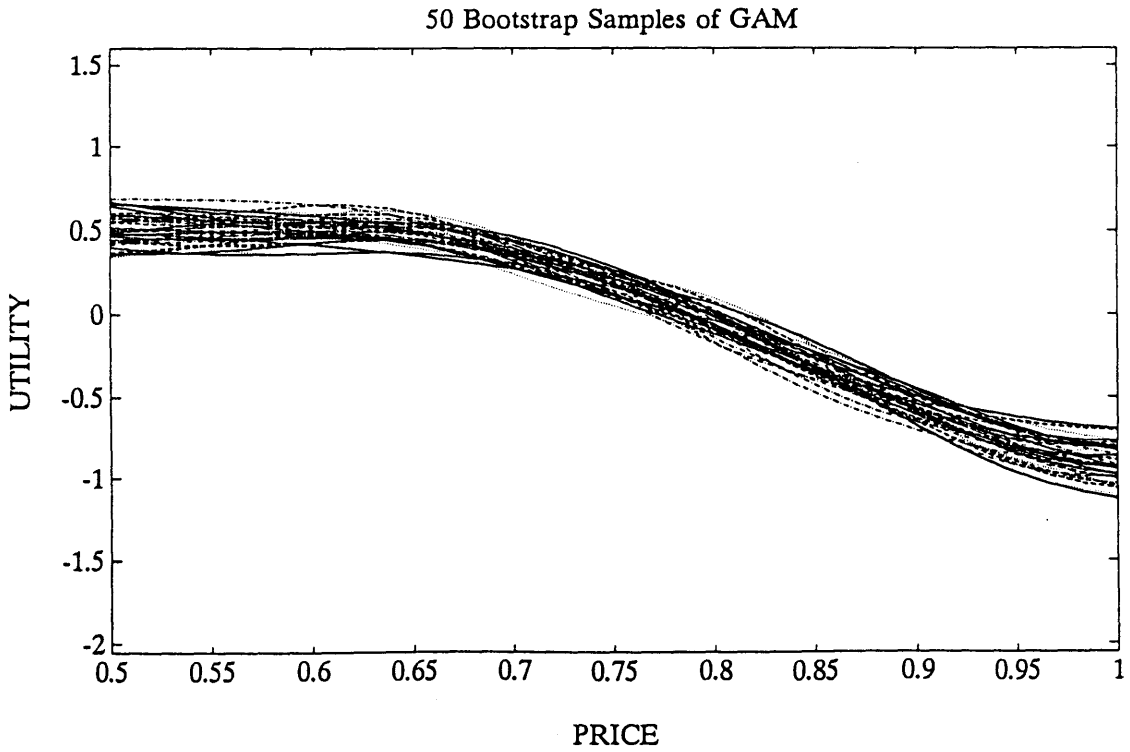
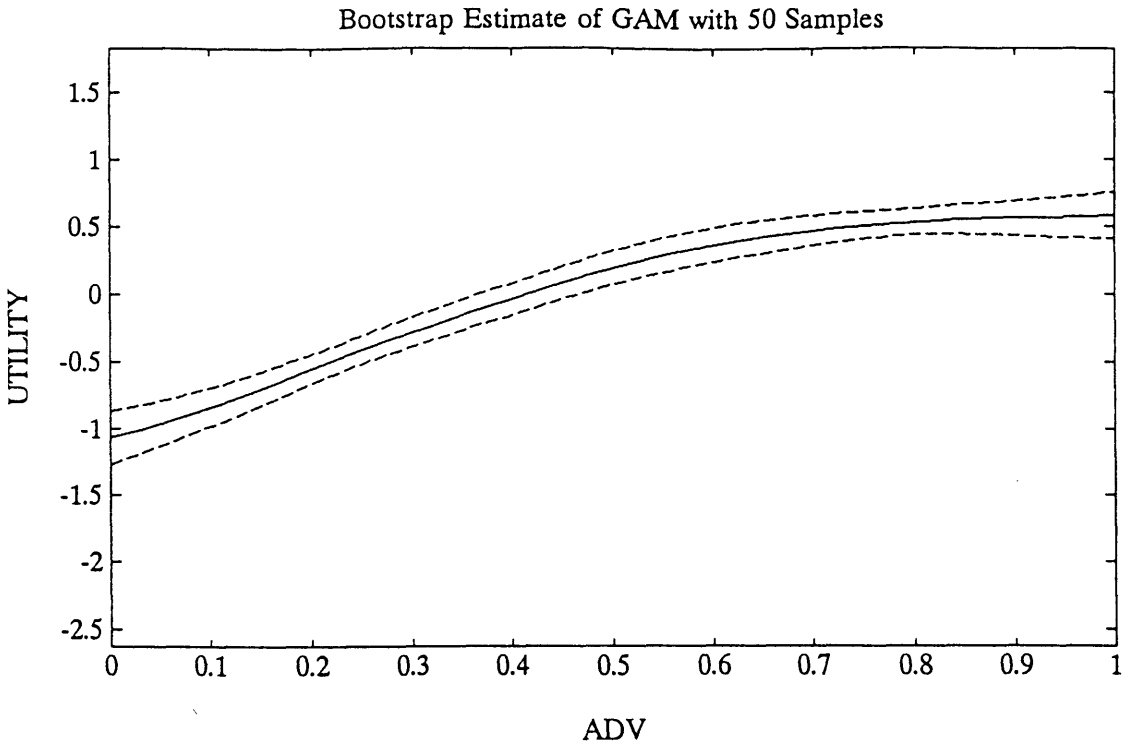


Figure C8: Bootstrap estimates of utility transformations with 50 samples



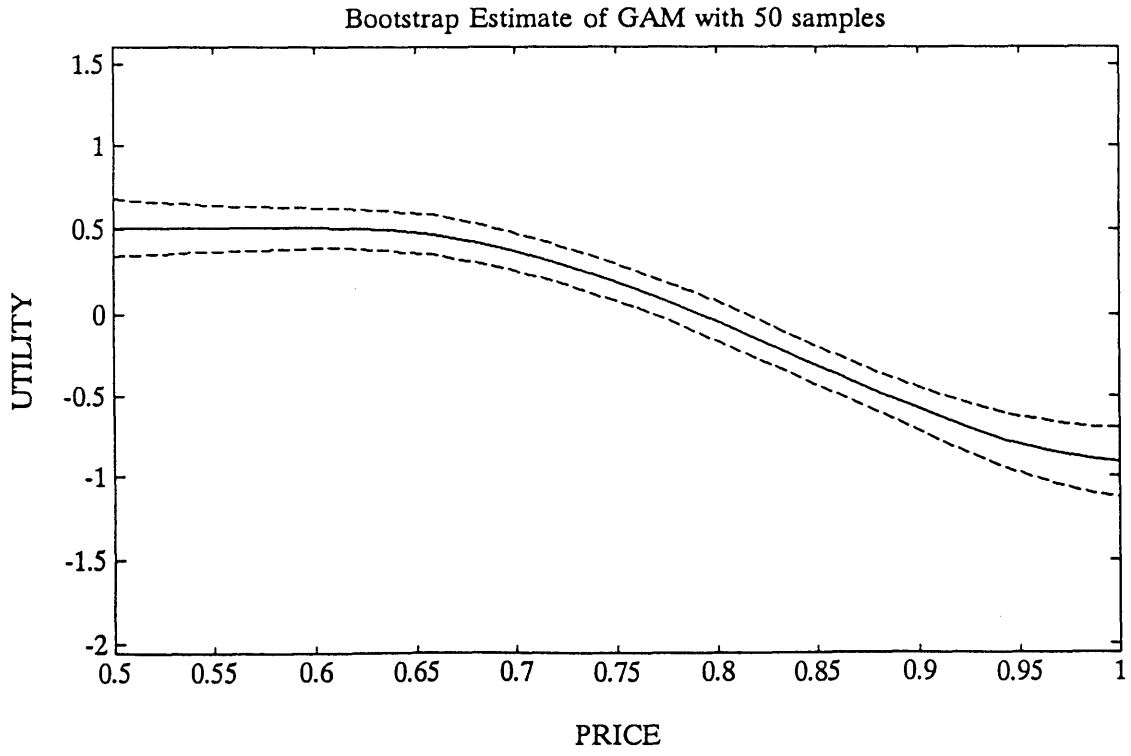


Figure C9: Approximate 95% pointwise confidence band for the utility transformations based on 50 bootstrap estimations

nonparametric regression, the main difficulty is not technicality or computation, but the amount of data required to obtain sufficiently reliable estimator, referred to as "curse of dimensionality". (Friedman and Stuetzle, 1981) Nevertheless, with close to 3000 observations in our simulation study, the benefit of using a two dimensional nonparametric function for variables between which interaction is highly suspected could outweigh the loss of some statistical reliability by uncovering hidden structures. Thus, we have attempted two dimensional kernel regression to compute the nonparametric utility transformation in advertising and price in both GAM and URM.

Figure C10 shows the result of GAM. Although the reconstruction resembles the original model in general, the additivity of the two variables specified in the utility function is not quite captured. The marginal effect of the advertising is larger for higher price and that of the price is larger for lower advertising. The issue needs further investigation, and a part of the reason must be attributed to the curse of dimensionality. Result of the two dimensional regression by URM is also presented in Figure C11. The smoothing constant (h=0.3) is chosen to be larger than that of the one dimensional case (h=0.2) because data points are more sparsely spread out in the two dimensional space. Again, the additivity of the simulation data is not quite captured.

To explore the additivity assumption of GAM and URM, an interaction between advertising and price is added to the original utility function and simulated choice data is generated. The additional interaction term is -6·adv·(price-0.5), which cancels the positive advertising effect as price is raised and reinforces the negative effect of price increase for higher advertising. As expected, the linear utility model estimates a weak positive coefficient for advertising, 0.656 with t=4.16, and a strong negative coefficient for price, -6.61 with t=-17.3.

Figure C12 shows the result of the GAM transformations. Note the small vertical scale in the advertising plot. t-value for advertising and price coefficients are 6.01 and 17.35 respectively, which are similar to those of the linear specification. In Figure C13, the two dimensional transformation is presented. Although it seems to resemble what the previous non-interaction data should have, a comparison with Figure C10 reveals that the positive effect of advertising is indeed dampened for higher price capturing the interaction. In terms of goodness-of-fit, the two dimensional GAM improves ρ^2 by 0.0323 over the additive linear model in comparison with 0.0157 for the additive GAM.

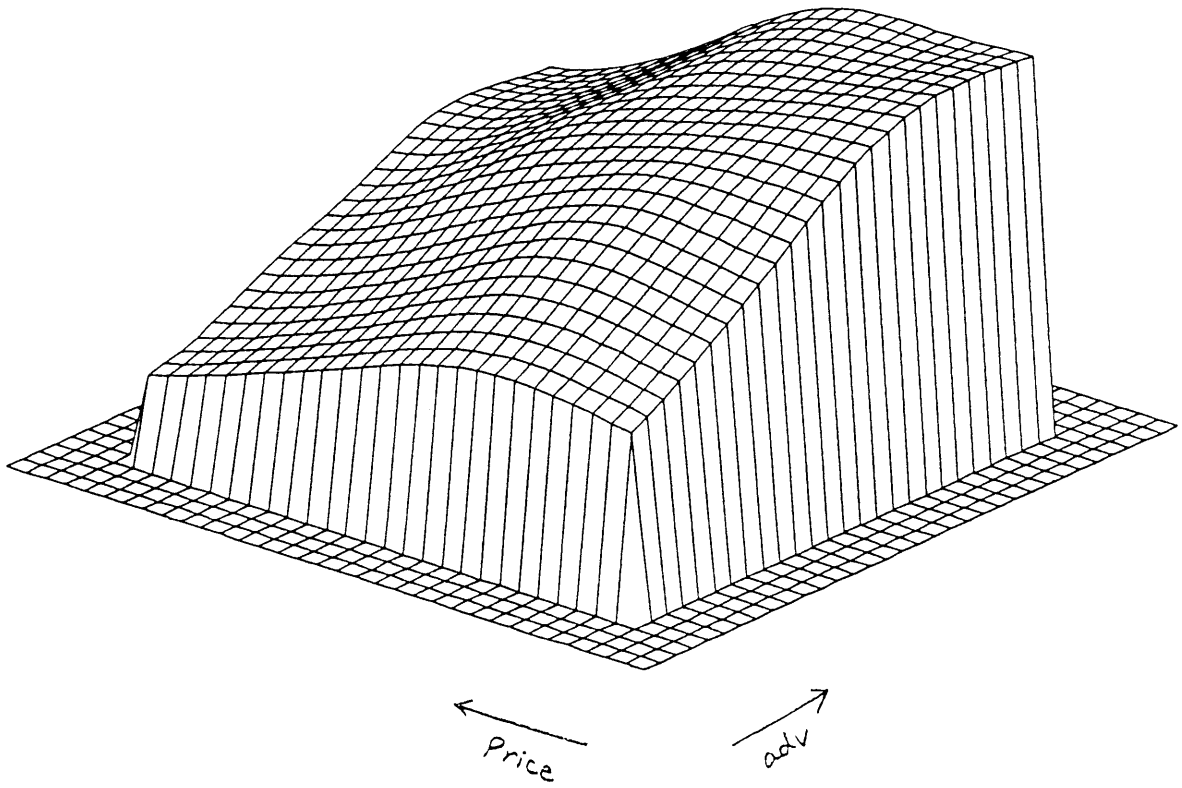


Figure C10: Two dimensional utility transformation by GAM on the simulation data

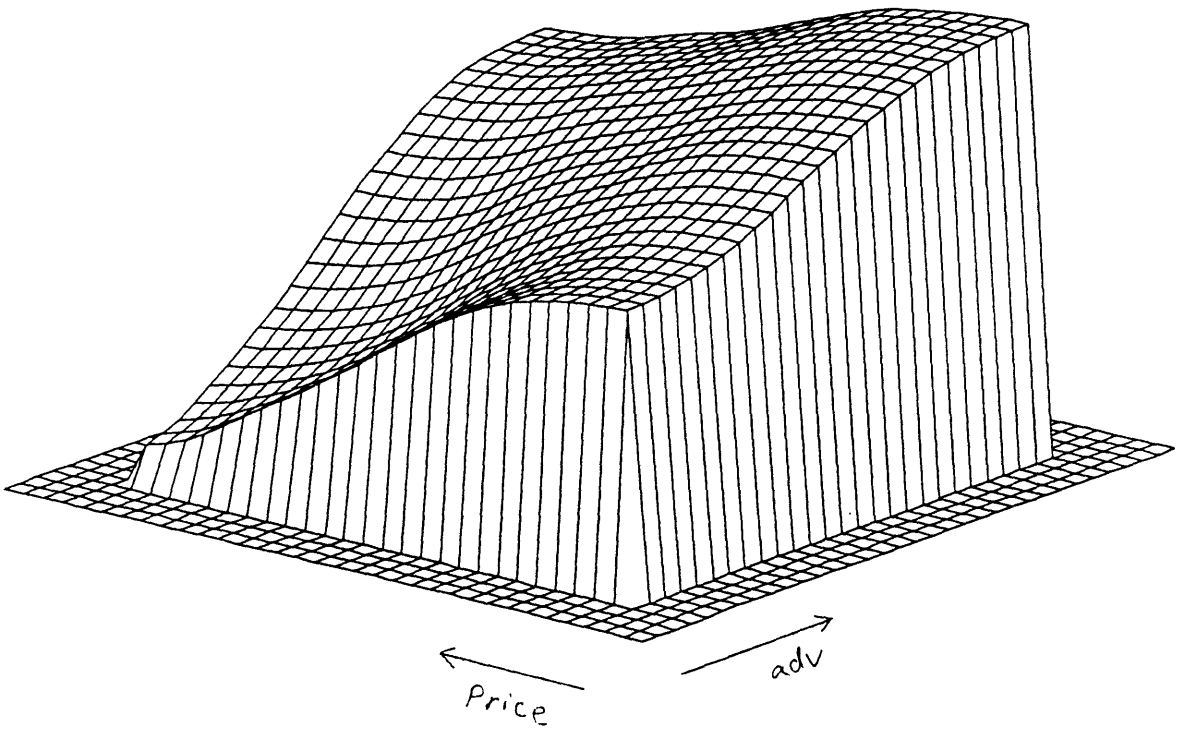
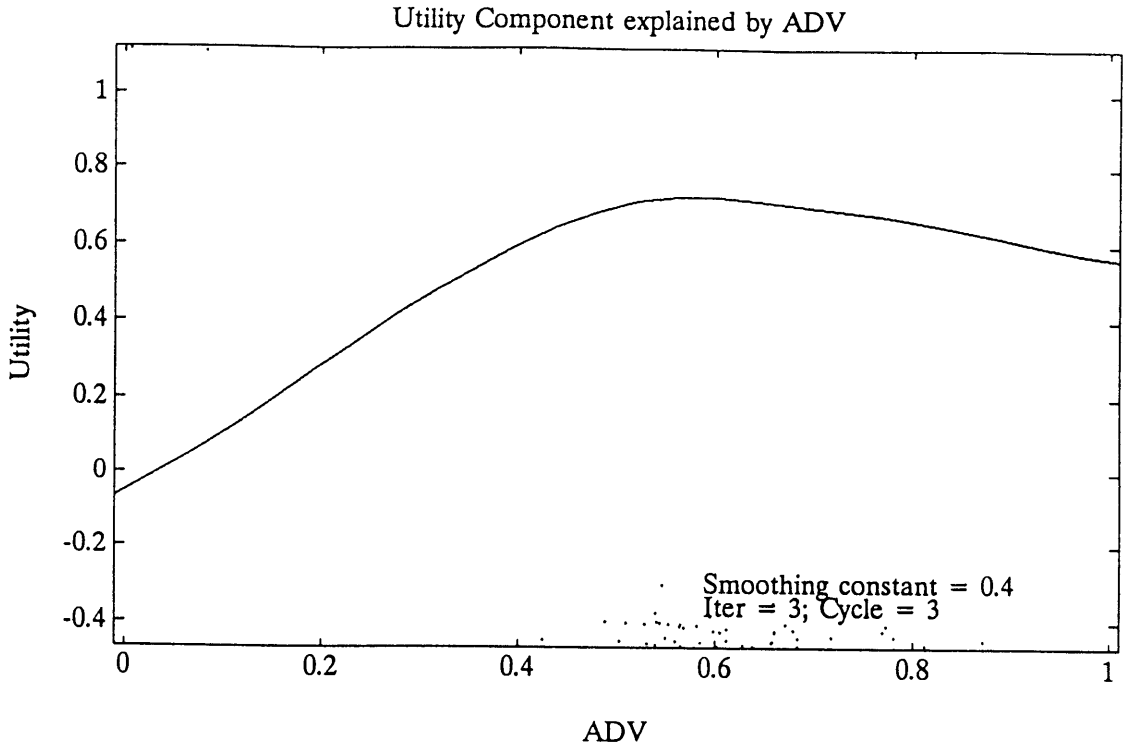


Figure C11: Two dimensional utility transformation by URM on the simulation data



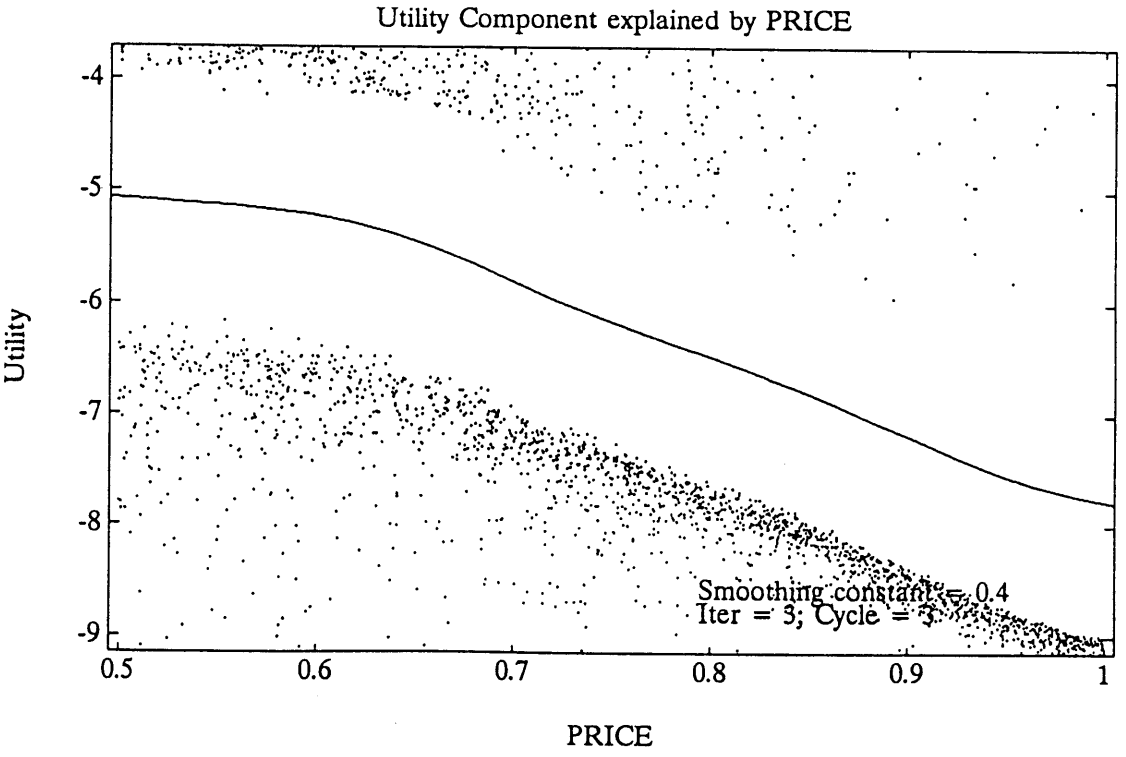


Figure C12: Utility transformations by GAM on the simulation data with advertising and price interaction

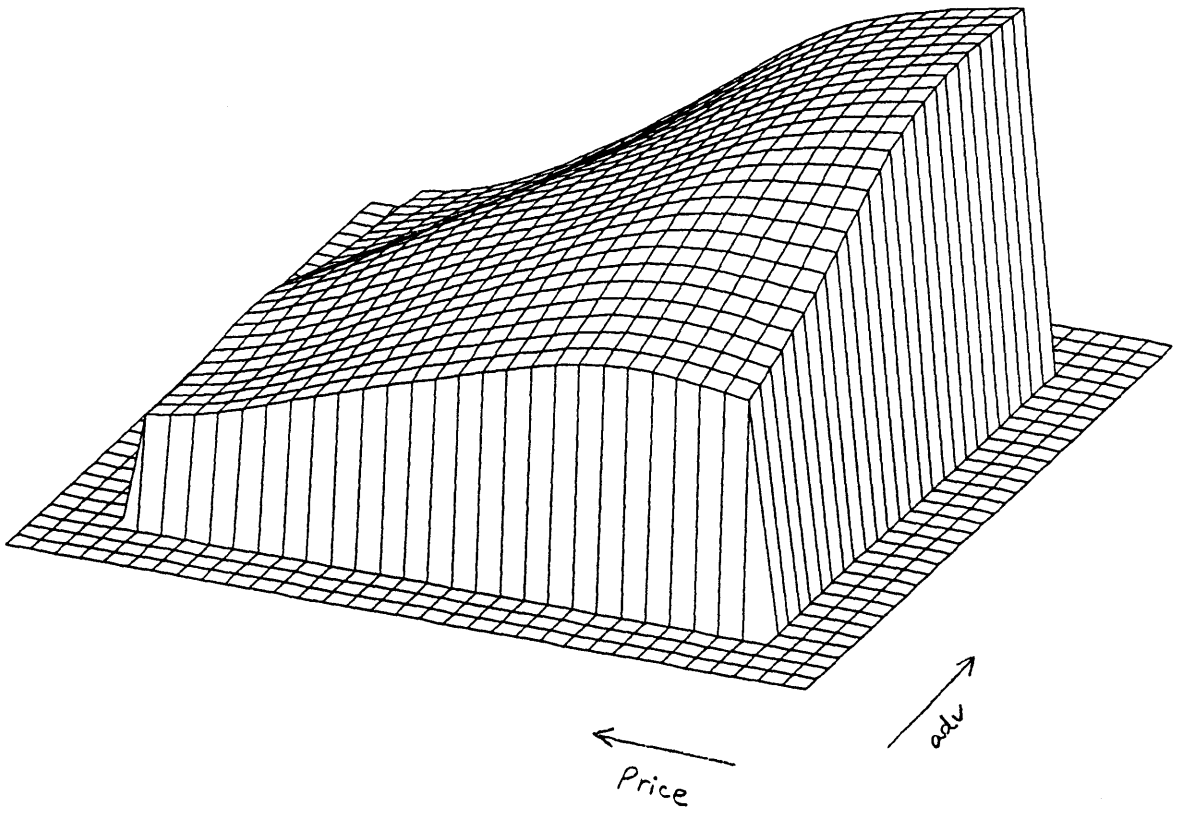


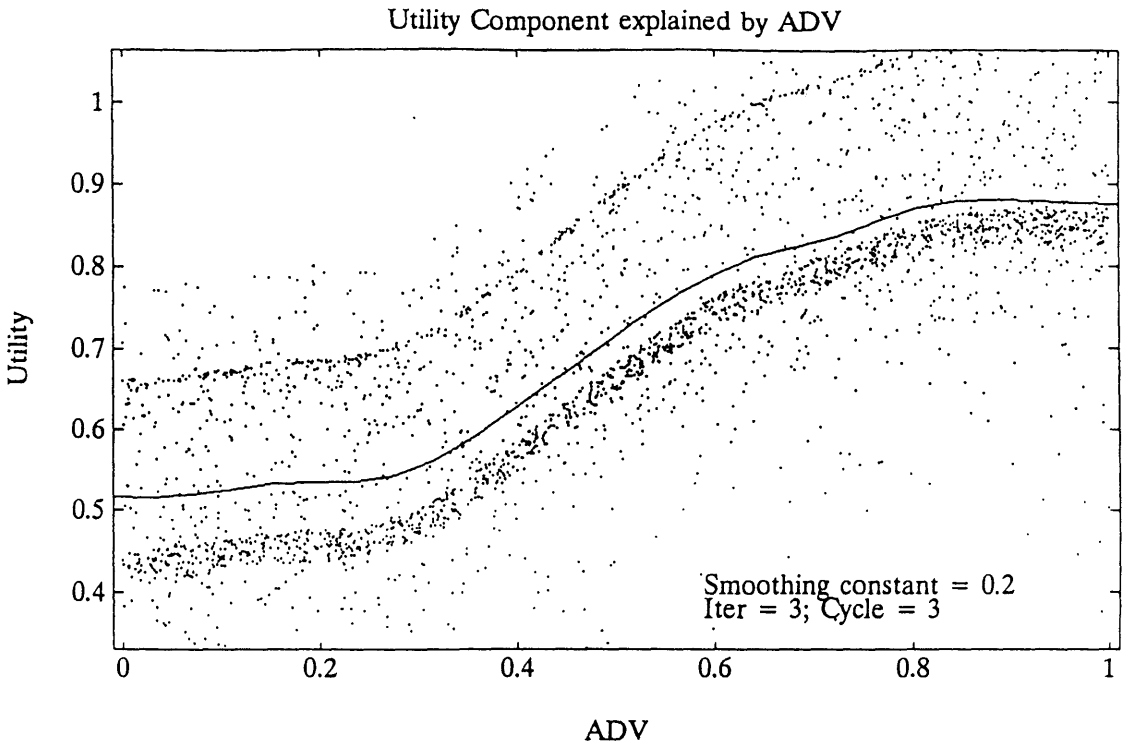
Figure C13: Two dimensional utility transformation by GAM on the simulation data with advertising and price interaction

Figure C14 shows the transformation obtained by URM from the same interacting data. The result is similar to the GAM in that the advertising, whose main and interaction term tend to cancel each other, has a much smaller scale than price where the interaction reinforces the main effect. Their t-values are 4.14 and 17.38 respectively, which are similar to the linear and GAM model. Figure C15 is a two dimensional transformations. It reproduces the decreasing utility for increase in advertising at price near $1 quite well. In terms of goodness-of-fit, the two dimensional URM improves ρ^2 to 0.21943 while the additive URM actually lowers ρ^2 to 0.21204, in comparison with 0.21379 achieved by the linear utility model. This study illustrates that when the additivity assumption of GAM and URM is violated, additive one dimensional transformation perform poorly. However, in URM one could at least visually observe a bad fit of the transformation from the partial utility residual plots and possibly conjecture the departure from the additivity assumption.

C.6 A Failing Example of the Extended GAM

The previous section demonstrates that the operational characteristics of the extended GAM surpasses those of URM in some aspects such as computational time and improvement in fit. As we learned, however, the method is less robust than URM towards various error distributions. In fact, the problem turned out to be much more serous than initially thought, and the GAM failed in some databases. Here I will report one such example on the Red Drink database used in Section 4.1 and discuss the reasons.

The extended GAM updates the utility by adding the "normalized residual", $(y-\mu)/\mu(1-\mu)$, to the current utility. Hence, if the predicted probability is close to either 0 or 1, the denominator becomes very small, while the numerator, the difference between the binary observation and the predicted probability, would remain non-zero even for a perfect model. This produces excessively large updated utility for the backfitting procedure. Figure C16 demonstrates the phenomenon in the Red Drink data just after the first iteration. Note the scale differences by a factor of 10^4 between the current and the updated utility. Estimated nonparametric utility functions at the end of the first iteration are presented in Figures C17, C18, and C19 for loyalty, price, and advertising respectively along with their rescaled plots so that all residuals are included. The algorithm could not proceed to iteration two due to a numerical precision error in computing MNL. The huge magnitude (~10,000) in the updated utility results in the counter-intuitive and insignificant utility functions.



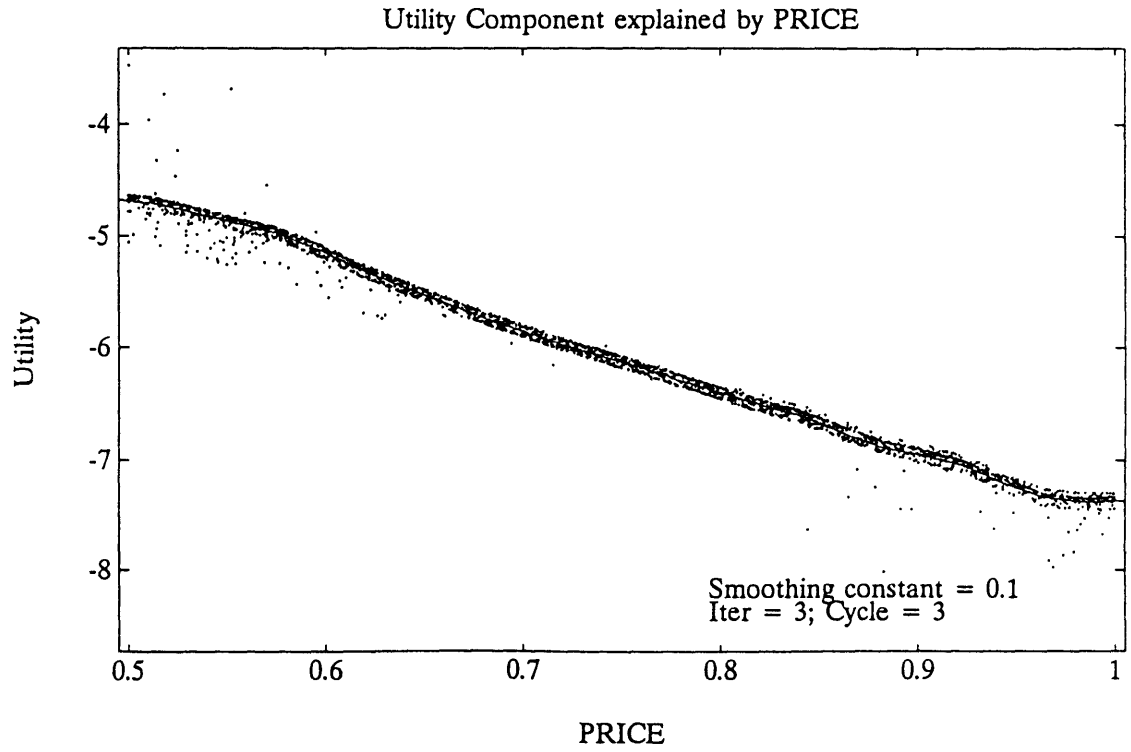


Figure C14: Utility transformations by URM on the simulation data with advertising and price interaction

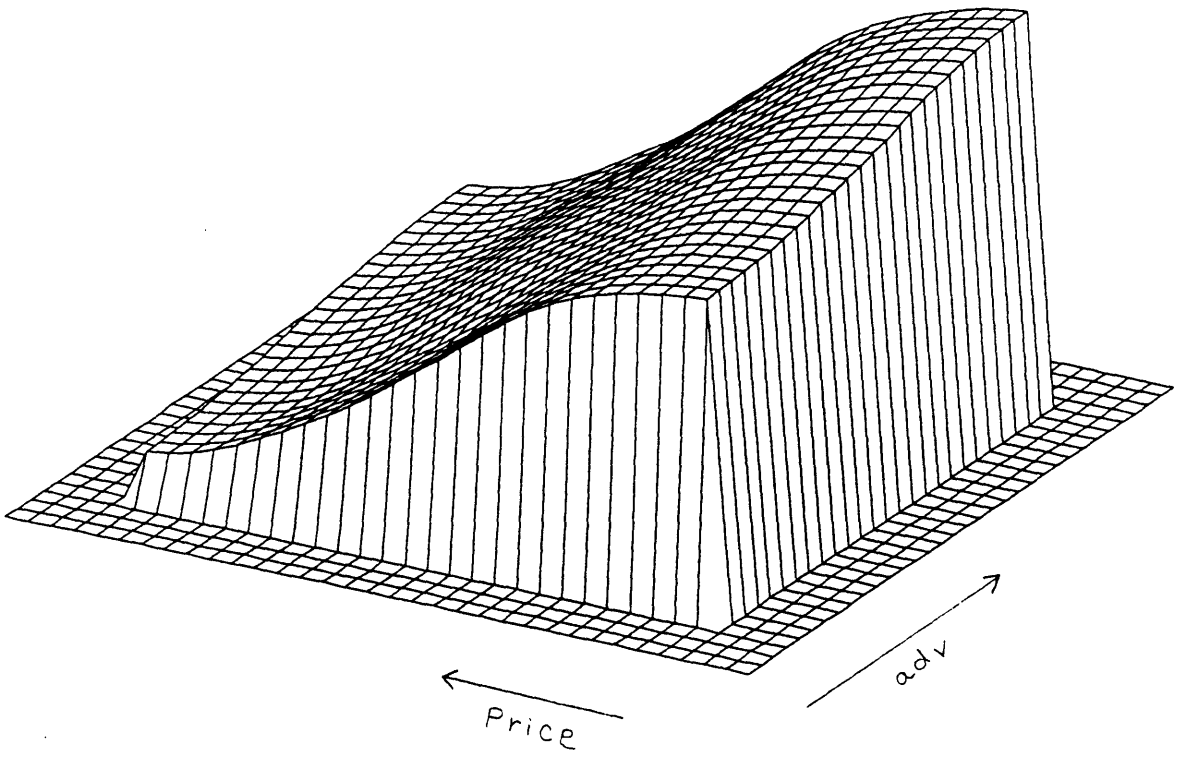


Figure C15: Two dimensional utility transformation by URM on the simulation data with advertising and price interaction

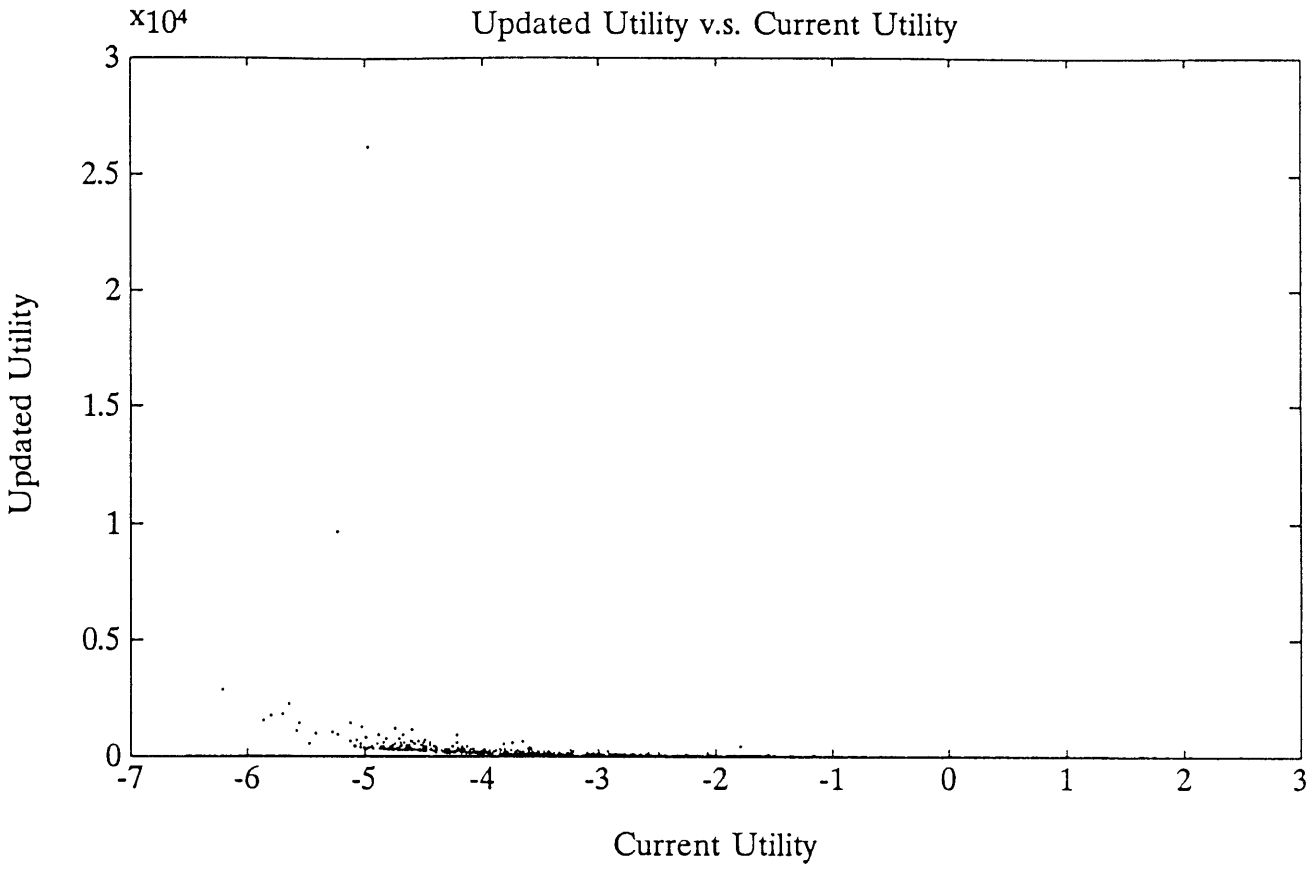


Figure C16: Updated utility v.s current utility in the Red Drink data

Because of the factor of $1/\mu(1-\mu)$ in updating the utility, new utility becomes enormous for a small value of μ as seen by the scale of the y-axis of $\times 10^4$.

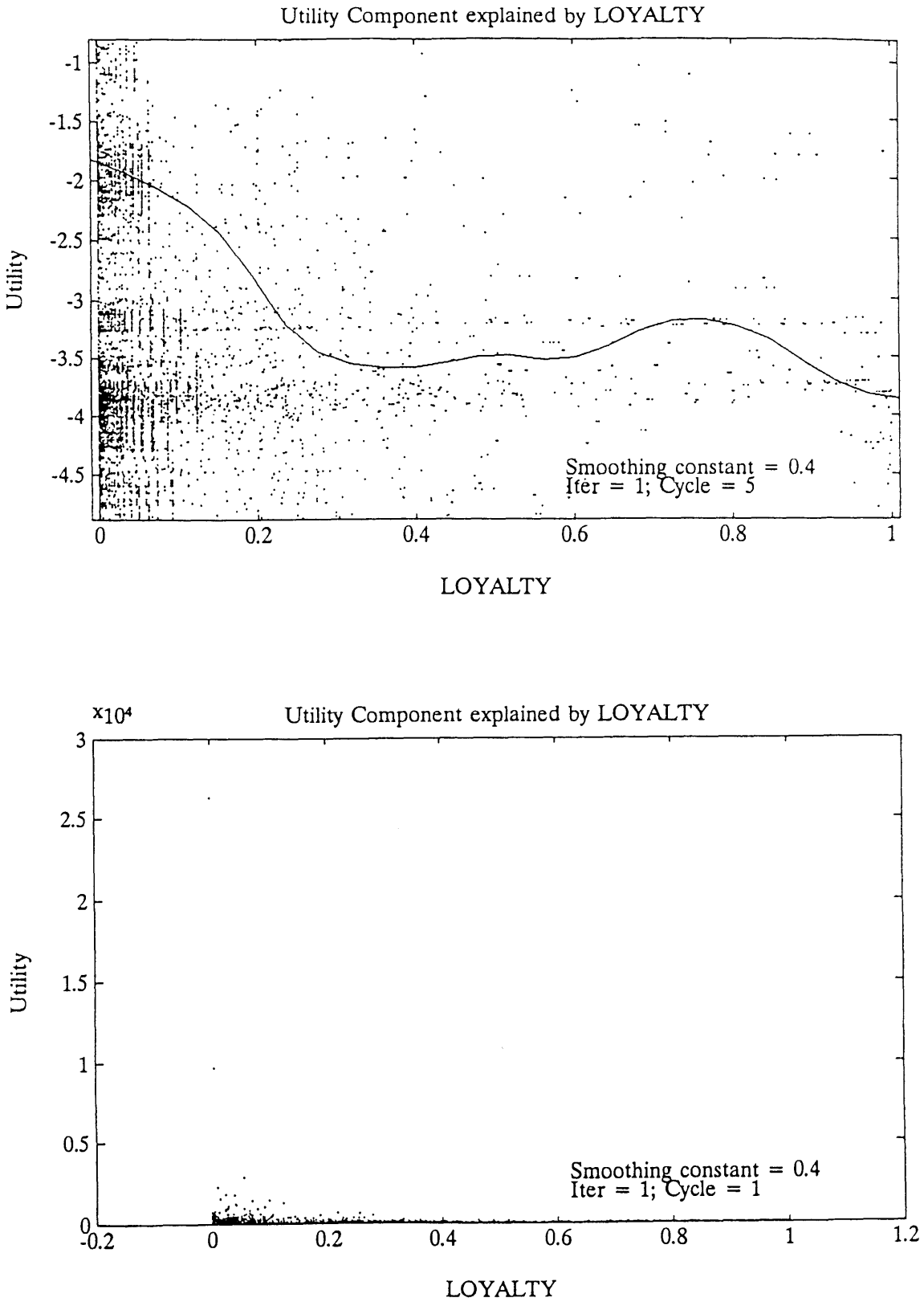


Figure C17: Additive nonparametric utility transformations of loyalty and its rescaled plot by the GAM in the Red Drink data

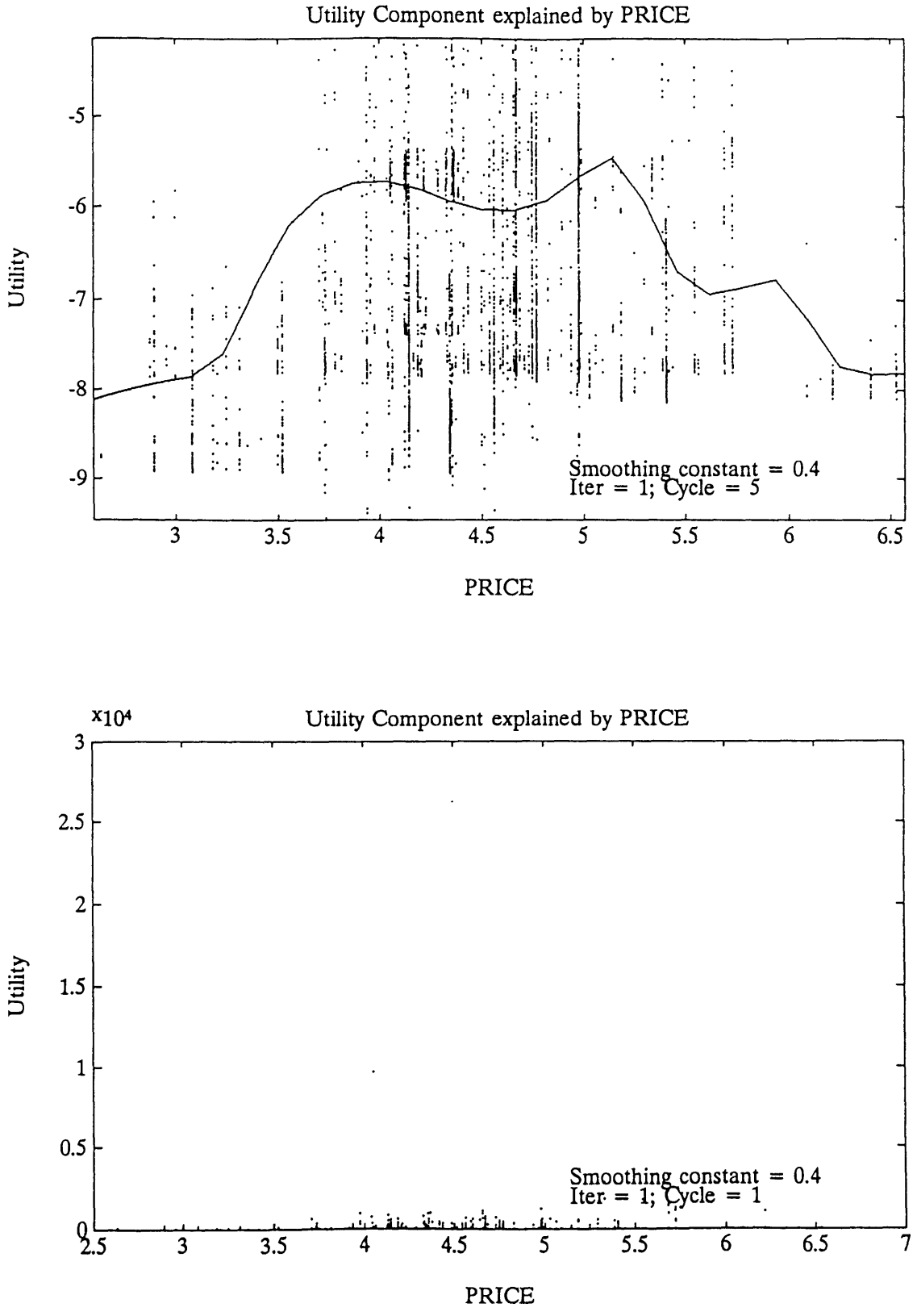
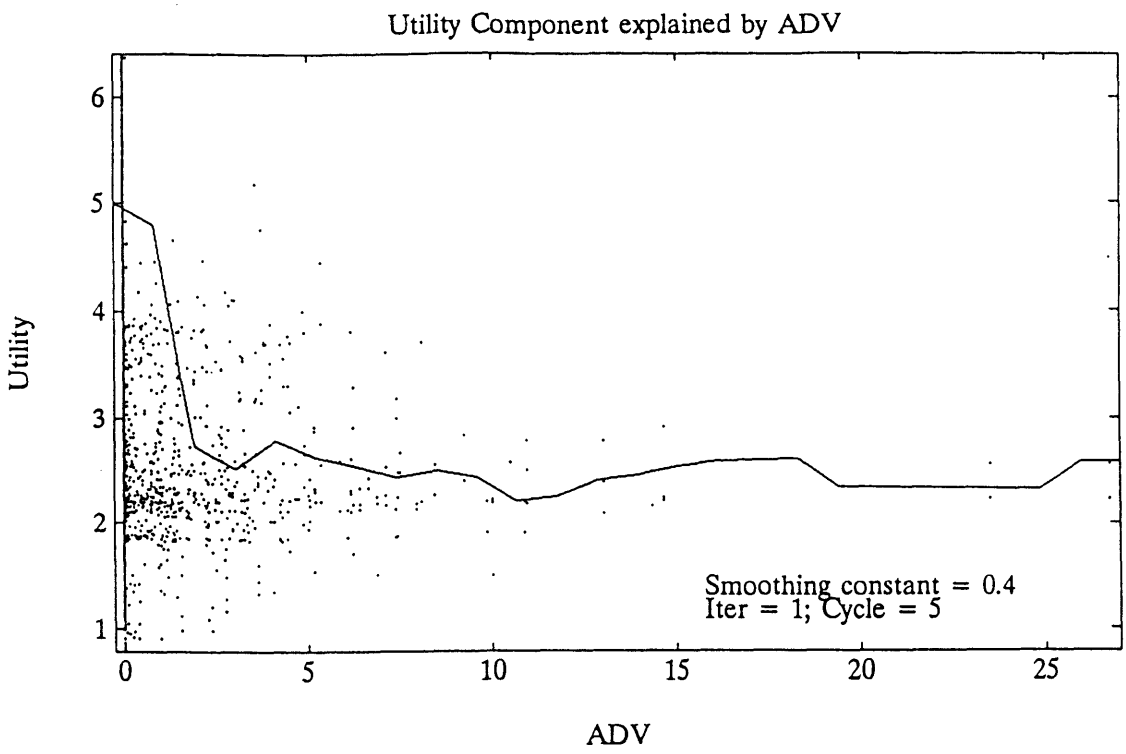


Figure C18: Additive nonparametric utility transformations of price and its rescaled plot by the GAM in the Red Drink data



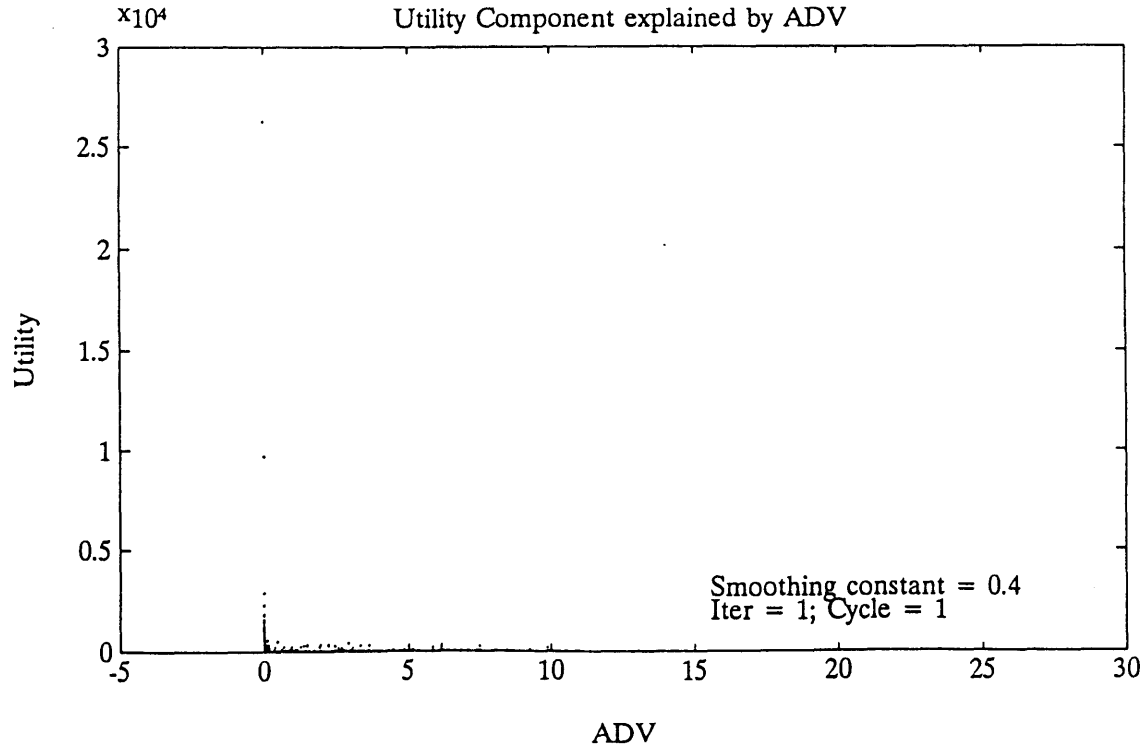


Figure C19: Additive nonparametric utility transformations of advertising and its rescaled plot by the GAM in the Red Drink data

Two methods are tried to overcome the difficulty: one by limiting the magnitude of $1/\mu(1-\mu)$, and the other by preprocessing the updated utility with repeated applications of running median in hope for eliminating outliers. The first remedy creates a cluster of points of the updated utility at the threshold value and distorts the whole information, while the latter failed to remove the outliers due to their overwhelming number.

This failure of the GAM is expected to be encountered rather frequently whenever extreme values are predicted for choice probabilities. Such a situation can occur when certain alternatives are either hardly chosen or chosen all the time, which is often the case in panel data. It could also happen by simply facing a moderately large number (~10) of alternatives.

APPENDIX D

Sufficient Statistic for Canonical Link in GLM

Following the notation of Appendix C, when $\theta=\eta$, the resulting link function is called a canonical link, and there exists a sufficient statistic for β. As an example, the canonical link for the Bernoulli distribution is a logit function since

$$\mu(\theta) = \frac{d\, b(\theta)}{d\,\theta} = \frac{d \log(1+e^{\theta})}{d\,\theta} = \frac{e^{\theta}}{1+e^{\theta}} = \frac{e^{\eta}}{1+e^{\eta}}$$

and thus $$\eta = g(\mu) = \log \frac{\mu}{1-\mu}.$$

Claim: The sufficient statistic for the canonical link is $\Sigma\, x_i\, y_i$.

Proof:

Assuming that $a(\phi)=1$ as usual, the likelihood function is obtained as

$$L = \exp\{\theta\, y - b(\theta)\},$$

and thus the loglikelihood function is

$$L = \theta\, y - b(\theta).$$

For the canonical link, $\theta=\eta=x'\beta$, where x and β are $p\times1$ column vectors, it is

$$L = y\, x'\beta - b(x'\beta) .$$

Because

$$\partial L/\partial\beta = x\, y - (\partial b/\partial\eta)\,(\partial\eta/\partial\beta) = x\,(y - \mu),$$

the first order condition for the loglikelihood with n observations is

(D1) $$\frac{\partial}{\partial\beta}\sum_{i=1}^{n} L_i = \sum_{i=1}^{n} x_i\,(y_i - \mu_i) \equiv 0 .$$

Define a static w (a $p\times1$ vector) with a probability density function $f_w(w;\beta)$ as

(D2) $$w \equiv \sum_{i=1}^{n} x_i\, y_i = \sum_{i=1}^{n} x_i\,\mu_i$$

using (D1).

To show that w is a sufficient statistic for β, it is adequate to demonstrate that

$$\frac{f(y_1;\beta)\, f(y_2;\beta) \ldots f(y_n;\beta)}{f_w(w;\beta)} \equiv A(\beta)$$

is independent of β. This is equivalent to

$$\frac{\partial A(\beta)}{\partial\beta} = 0 \quad \text{or} \quad \frac{\partial \log[A(\beta)]}{\partial\beta} = 0 .$$

But,

$$\begin{aligned}\frac{\partial \log[A(\beta)]}{\partial\beta} &= \frac{\partial}{\partial\beta}\left[\sum_i L_i - \log f_w(w;\beta)\right] \\ &= \frac{\partial}{\partial\beta}\sum_i L_i - \frac{\partial}{\partial\beta}\log f_w(w;\beta) \\ &= 0\end{aligned}$$

because the first term is 0 by the first order condition (D1), and so is the second term as w is independent of β by the first identity of (D2).

Q.E.D.

APPENDIX E

Sketch of the Derivation for the Local Scoring Algorithm

[1] The first order condition for the expected loglikelihood function is

$$E\left(\frac{dL}{d\eta}\,\middle|\, x\right)_{\eta=\hat{\eta}(x)} = 0$$

assuming that the integration and differentiation can be exchanged. To solve for $\hat{\eta}(x)$ iteratively, a new estimate of $\eta(x)$, $\eta_n(x)$, is obtained by applying the Newton-Raphson method as

(E1) $$\eta_n(x) = \eta(x) - \frac{E(dL/d\eta \mid x)}{E(d^2L/d\eta^2 \mid x)}$$

For the exponentially family distributions of (C1), it can be shown that

$$\frac{dL}{d\eta} = \frac{dL}{d\theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta} = (y-\mu)\,v^{-1}\frac{d\mu}{d\eta}$$

$$E\left(\frac{d^2L}{d\eta^2}\,\middle|\, x\right) = -\left(\frac{d\mu}{d\eta}\right)^2 v^{-1}$$

after some manipulation analogous to McCullagh and Nelder (1983, p41). Substituting these in (E1) results in

(E2) $$z = E\left[\hat{\eta} + (y-\hat{\mu})\left(\frac{d\eta}{d\mu}\right)_{\mu=\hat{\mu}}\,\middle|\, x\right]$$

Note that the adjusted dependent variable for GAM shown in (E2) is identical to its counterpart for GLM of (C2) except for the expectation. For GAM with the additive nonparametric predictors, the conditional expectation operator in (E2) is replaced by the backfitting nonparametric regression with the weights

(E3) $$w^{-1} = \text{var}(\eta_n) = v\left(\frac{d\eta}{d\mu}\right)^2_{\mu=\hat{\mu}}$$

instead of the weighted least square regression in GLM.

The Utility Residual Method as a Diagnostic Tool for Building a Nested Logit Marketing Mix Model from Single Source Data

Makoto Abe

Operations Research Center
M.I.T.
Cambridge, MA 02139 USA

M. I. T. Doctoral Dissertation, Part III

June 1991

OVERVIEW

Marketing managers in packaged goods companies now have access to great quantities of scanner panel data which allow household level modeling, which is particularly useful for understanding marketing effects on the final customer and providing a deeper grasp of market responses. In a household level marketing mix model, a purchase process is considered to consist of three inter-dependent elements: namely, purchase incidence, brand choice, and quantity selection. Although many disaggregate models exist for each element, less effort has been devoted to combining them together to build full scale marketing mix models directed at the marketing manager's brand planning problem.

In this regard, the nested logit formulation is introduced in the current work because it is highly integrated, easily handles quantity purchases, and is driven by shopping trips. The scanner data used in this study is an IRI single source database which contains store records, panel purchase and shopping trip data, as well as household TV advertising exposures monitored by TV meters. Explanatory variables for the model are constructed based on behavioral considerations and take into account carry-over and forgetting of advertising exposures, dynamic change and heterogeneity in household consumption rate, and inventory.

The variables are studied by the utility residual method described in Part II --- a method to obtain an additive nonparametric utility function in logit models --- to infer appropriate parametric covariate transformations. Graphical diagnostics in logit models are found to be very useful for identifying influential points, outliers, and heterogeneous segments. The parameters are estimated in a calibration sample, their marketing implications are discussed, and the model is cross-validated in a holdout sample.

1. INTRODUCTION

1.1 Motivation

Marketing managers in packaged goods companies now have access to great quantities of scanner data which contain not only sales but also information on price, promotional activities and, increasingly, advertising. Of particular interest is the growing amount of household panel data. Although many academic papers have made use of this data, there has been a surprising lack of full scale marketing mix models based on the data and directed at the marketing manager's brand planning problem. While an aggregate level marketing mix model which links sales and/or share to control variables of a firm is valuable for planning and strategy (Little 1975, Abraham & Lodish 1987, Blattberg & Levin 1987), it does not usually suggest the underlying buyer behavior and how specific marketing implementations affect consumer purchase decisions. Household level analysis holds the promise of avoiding certain aggregation biases due to household heterogeneity and seems rich for understanding buyer behavior.

In a household level marketing mix model, relationships are sought between marketing variables and inter-dependent elements of a purchase process such as purchase incidence, interpurchase time, brand choice, and quantity selection. Although many disaggregate models exist for individual elements, only a few consider them together. Most of these studies combine separate models, which work largely independently for each element. Hence, there is potential advantage in an integrated marketing mix model at a household level to describe the underlying consumer process. Such models will allow the researcher to obtain better understanding of dynamic marketing phenomena as well as managers to act more properly on strategic and tactic issues.

Effectiveness of sales promotions and advertising can be further enhanced by appreciating market response in detail. Information contained in panel data in conjunction with behavioral theories could fine tune the final execution of marketing activities by examining many issues involved in market response including various possible nonlinearities, asymmetries, and interactions. In addition to numerous behavioral studies conducted in laboratory settings, there exists increasing number of field studies utilizing panel data to explore market response. Reference pricing (Winer 1986, Kalwani et. al. 1989, Gurumurthy & Little 1989) and reference promotion studies (Lattin & Bucklin 1989) shed light on the long term impact of frequent price

cut and promotions in repeat purchasing. Advertising exposure response has been studied by Tellis (1988), and Kenetker, Weinberg, & Weiss (1989).

The current research develops a household level marketing mix model which integrates the three elements of a buying decision --- namely, purchase incidence (when), brand choice (what), and quantity selection (which size and how many) --- in integrated framework with special attention to nonlinear market responses and their long term effects. The model is formulated based on the nested logit model (Guadagni & Little 1987) with various carry-over variables, whose responses are diagnosed nonparametrically by the utility residual method. Unlike Part II, URM is used conservatively for graphical inspection rather than directly incorporating the resulting nonparametric utility transformations into the model. In the current work, the diagnostic capability of URM is exploited.

1.2 A Brief Review of the Literature

Table 1 summarizes a number of relevant modeling studies of the four elements, purchase incidence (PI), interpurchase timing (IT), brand choice (BC), and quantity selection (PQ) in a buyer decision process. The list is created with a particular emphasis on market response of sales promotions and advertising. It is by no means an exhaustive list.

The well-known negative binomial distributed (NBD) interpurchase timing models aggregate over heterogeneous consumers by assuming gamma distributed purchase rate. (Ehrenberg 1972) Such aggregation approaches to the brand choice are mostly based on stochastic processes such as zeroth order, Markov, and linear learning (Kuehn 1962, Lilien 1974) models. One of the first multiple element purchase process was proposed by Jeuland, Bass, & Wright (1980) using NBD interpurchase timing compounded with the multinomial Dirichlet choice model. Their model assumes independence of the two elements and marketing variables are not introduced. Neslin, Henderson, & Quelch (1985) and Wagner & Taudes (1986) integrate interpurchase timing and purchase quantity which take into account the effect of marketing mix variables by aggregating panel data. The former studies promotional purchase acceleration by a simultaneous regression equation technique and finds that the acceleration is more likely to be attributed to increased purchase quantity than shortened interpurchase time in both bathroom tissue and coffee product category. The latter incorporates marketing mix variables and gamma distributed heterogeneity to the Poisson incidence model coupled with the zeroth order generalized

Table 1

Household level marketing mix models focusing on market response issues

Article	A/D	S/E	H	N	PI	IT	BC	PQ	DEP
Hauser & Wisniewski 1982	A	S(1)	Y	Y	yes	-	yes	-	Y
Carpenter & Lerman 1985	A	S(1)	N	N	-	-	yes	-	NA
Neslin et. al. 1985	A	E	N	N	-	yes	-	yes	Y
Wagner & Taudes 1986	A	S/E	Y	Y	NBD	-	yes	-	N
Jones & Zufryden 1980	A/D	S/E	Y	N	NBD	-	MNL	-	N
Zufryden 1987	A/D	S	Y	N	yes	-	-	-	NA
Krishnamurthi & Raj 1985	A	E	Y	N	-	-		yes	NA
Guadagni & Little 1983	D	E	Y	Y	-	-	MNL	-	NA
Lattin 1987	D	E	Y	Y	-	-	MNL	-	NA
Lattin & Bucklin 1989	D	E	Y	Y	-	-	MNL	-	NA
Kalwani et. al. 1989	D	E	Y	Y	-	-	MNL	-	NA
Kanetkar et. al. 1989	D	E	Y	Y	-	-	MNL	-	NA
Gurumurthy & Little 1989	D	E	Y	Y	-	-	MNL	-	NA
Guadagni & Little 1987	D	E	Y	Y	NL	-	NL	NL	Y
Tellis 1988	D/A	E	Y	Y	-	-	MNL	CR	Y
Krishnamurthi & Raj 1988	D	E	Y	Y	-	-	MNL	CR	Y
Gupta 1988	D	S/E	Y	Y	-	Erl-2	MNL	cumL	N
Pedrick & Zufryden 1990	A/D	S/E	Y	Y	NBD	-	MNL	-	N

NOTATION: A/D = Aggregate or Disaggregate purchase occasion approach
S/E = Stochastic or Econometric model
H = does it take into account of Heterogeneity of the households?
N = does it take into account of Nonstationarity?
PI = Purchase Incidence
IT = Interpurchase Timing
BC = Brand Choice
PQ = Purchase Quantity
DEP = does it consider inter-dependency among elements, PI, IT, BC, and PQ?

S(x) = Stochastic model of order x
NBD = Negative Binomial Distribution model
MNL = Multinomial Logit Model
NL = Nested Logit model
Erl-2 = Erlang 2 distribution model
CR = Censored Regression
cumL = Cumulative Logit model
NA = Not Applicable

Dirichlet distribution brand choice process (Jeuland 1978), which is called Polya process, and demonstrates a good time series tracking to the detergent data.

There have been increasingly many disaggregate studies with further complexity in an effort to combine more than one element of purchase process. Guadagni & Little (1987) integrate brand choice, quantity selection, and purchase incidence driven by shopping trips, in the nested logit formulation. Both Tellis (1988) and Krishnamurthi & Raj (1988) combine the multinomial logit brand choice model with purchase quantity in a regression framework using limited dependent variable techniques (Maddala 1983, Amemiya 1985). Their substantive focuses differ in that the former investigates the effect of interaction between advertising exposure and brand loyalty on brand choice and purchase quantity, while the latter considers price elasticities. Gupta (1988) merges the three independent elements, interpurchase timing by Erlang-2 distribution, brand choice by logit, and purchase quantity by cumulative logit. He concludes that much of the promotional effect appears on brand switching with small effect on purchase timing and negligible effect on stockpiling (increased purchase quantity) in coffee category, contrary to the earlier finding by Neslin et. al. (1985). Pedrick & Zufryden (1990) examine the impact of advertising media plan, in particular depth and frequency, by the logit brand choice model compounded with the Poisson purchase incidence assuming their independence.

1.3 Issues in Multiple Element Purchase Process Models

There are three major issues to be considered when combining more than one element of the purchase process.

1.3.1 Inter-dependency among the elements

By far the most complete combined models to date are those by Guadagni & Little (1987)[10] and Gupta (1988), in which all three elements, purchase incidence or interpurchase timing, brand choice, and purchase quantity, are taken into account. However, there is a difference between the two in the way to combine the elements. The nested logit formulation, as will be discussed

---

[10] Although they did not emphasize the quantity selection other than size choice (how many additional units to buy), their model can be easily extended to model a more sophisticated purchase quantity scheme by introducing interaction terms with the "first purchase opportunity within trip" dummy variable, if enough observations are made on multiple-unit purchase.

in section 2.1, allows the decision of purchase incidence and purchase quantity to depend on brand choice via "the category attractiveness" which refers to the expected maximum utility of all available brands at the shopping trip. Therefore, joint probability of purchase incidence (PI), brand choice (BC), and purchase quantity (PQ) under the influence of explanatory variables, X, is expressed as

$$P(PI, BC, PQ \mid X) = P(PQ \mid PI, BC, X)\, P(PI \mid BC, X)\, P(BC \mid X)\,.$$

While Gupta's model contains three independently working modules so that

$$P(PI, BC, PQ \mid X) = P(PI \mid X)\, P(BC \mid X)\, P(PQ \mid X)\,.$$

The former seems more appealing. The following example will illustrate some of the dependency issues.

> *Suppose Mrs. Logit went for grocery shopping but was not planning to buy any ground coffee because she had plenty in stock at home. (This implies a very low category purchase probability if the purchase incidence model contains a household inventory variable.) While she was in the store, a huge stack of coffee cans with a big sign saying "Today Only, Half Price on Miniwell House Coffee" caught her attention. The Miniwell brand has one of the lowest market shares, and Mrs. Logit has never bought it before. But considering the low price, she decided to buy the coffee.*

In both methods, Guadagni & Little and Gupta, the brand choice model will predict a choice of Miniwell brand with high probability. The independent model is likely to predict a low category purchase probability, however, because the effect of the Miniwell promotional price cut is weakened by the use of share weighted category price and promotional variables in the interpurchase timing model. In other words, the effect of promotions may be underestimated in category purchase and quantity model by the formulation. In contrast, the nested logit will pick up the promotional effect via category attractiveness (because Miniwell brand results in the maximum utility for that trip) and thus boost the category purchase probability. As this "thought experiment" indicates, the independence assumption among the purchase process elements seems less desirable than the full conditioning.

The formulations of Tellis (1988) and Krishnamurthi & Raj (1988) are basically the same. They both capture the dependency between brand choice and purchase quantity by establishing brand specific quantity equations. Furthermore, their quantity model is censored regression which interacts with the brand choice model. Pedrick & Zufryden (1990) compound MNL

brand choice into NBD purchase incidence by using the predicted brand choice probability as the purchase rate parameter of the Poisson incidence model under their independence assumption, which is a quite common technique in stochastic purchase models. (Jeuland et. al. 1980)

1.3.2 Purchase incidence v.s. Interpurchase timing

Wheat & Morrison (1990) argue that purchase incidence is a better way to model category purchase than interpurchase timing for two reasons. First, category purchase decisions are, by and large, driven by shopping trips, implying that schedules of many people are not flexible enough to allow irregular shopping trips. Second, there exists a possibility of selection bias caused by truncating observations of very long interpurchase times from a study sample.

In addition, an accurate estimation of the relation between interpurchase timing and intensity of promotions is difficult due to confounding with the timing of promotions observed in the data. In other words, the extent to which interpurchase time is shortened is affected not only by the effectiveness of the promotion but also when the promotion takes place --- a factor which is out of our design control. Such a difficulty does not arise in an analysis of the relationship between purchase incidence probability and promotion intensity.

It is also technically burdensome to incorporate time varying marketing mix variables in interpurchase timing models due to commonly practiced weekly promotions and price changes. (Gupta 1991)

1.3.3 Discrete nature of purchase quantity

Both Tellis and Krishnamurthi & Raj treat purchase quantity as a continuous variable and use regression in their models. The approach could result in biased estimates for observed quantities which are integral as 1 unit, 2 units, or multiples of package sizes such as 16oz and 32oz. Another complication here is how to define marketing variables for a brand which consists of different sizes. Predictor variables such as price and promotions are generally size specific. Thus, setting them as share or sales weighted values --- the method employed by many studies --- smooths out much variations in these variables and discards important information in the data. In contrast, the nested logit model, which differentiates alternatives by brand and size and views each additional unit purchase as a separate purchase occasion, poses no problem.

1.4 Model

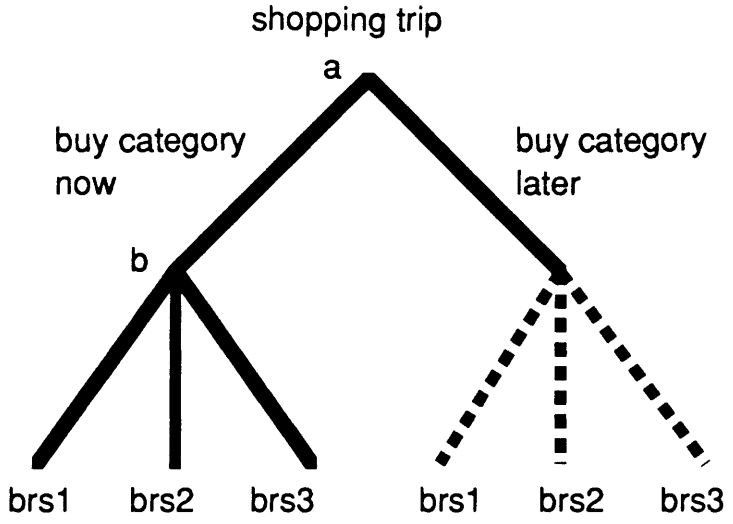
With these factors taken into account for the combined element approach, the decision has been made to base the model on the nested logit formulation. The nested logit is appealing from theoretical point of view since it is derived from the utility maximization consumer theory unlike many regression and stochastic models.

This paper is organized as follows. Section 2 illustrates the methodology to be used in our model by briefly reviewing the nested logit model of Guadagni & Little (1987). Then the utility residual method (URM), a technique used for diagnosing the nonlinearity in the utility function, is introduced. Section 3 describes the Red Drink single source data for the study. Then, the brand choice and category purchase models are discussed and their calibration results are presented in section 4 and 5 respectively. Finally, section 6 summarizes the study.

2. METHODOLOGY

2.1 Nested Logit Model

In the nested logit (Ben-Akiva & Lerman 1985) as adapted by Guadagni & Little (1987), the consumer purchase may be thought of as a two stage-process of category purchase followed by brand choice in an integrated framework. (However, as noted by Guadagni & Little, theory does not require that the actual customer decision making is in two stages.)



Nested logit model for consumer purchase process

The key concept of the nested logit hierarchy is an interaction between two stages, the upper branch and its subtree, where the former choice depends on the utilities associated with all alternatives in the subtree in addition to its own branch characteristics. In our case, *buy category now* is influenced by attributes of stage *a* decision as well as utilities of choosing each brandsizes given *buy category now* is selected. Hence, choice probability of the consumer's ultimate alternative (a,b) can be expressed as $P(a,b \mid x) = P(a \mid b,x)\, P(b \mid x)$ rather than $P(a,b \mid x) = P(a \mid x)\, P(b \mid x)$ which implies independence of the two stages.

Nested logit postulates that the utility of alternative (a,b) is expressed as

$$u_{ab} = w_a + v_b + v_{ab} + e_a + e_{ab}$$

by assuming the variance of e_b to be negligible compared to that of e_a. Then,

$$P(b \mid a) = \frac{e^{V_b + V_{ab}}}{\sum_{k \text{ under } a} e^{V_k + V_{ak}}} = \frac{e^{V_b + V_{ab}}}{e^{W'_a}}$$

and

(1) $$P(a \mid b) = \frac{e^{(w_a + w'_a)\mu_a}}{\sum_{k \subset A} e^{(w_k + w'_k)\mu_a}}$$

where w'_a = systematic component of the maximum utility of the brandsize alternatives b's that involve *a*

= expected maximum utility of the brandsize alternatives b's under branch *a*

w'_a is the interaction between upper branch and its subtree and referred to as a measure of the "inclusive value" (Ben-Akiva & Lerman 1985). In our context, we call it "category attractiveness" for its intuitive interpretation. For example, if a consumer's favorite brandsize is on promotion, not only its choice probability increases but so does the probability of the purchase incidence via increase in the category attractiveness. v_{ab} is not necessary in our case because no brandsize choice takes place when *buy category latter* is chosen. Thus, $v_{ab} = 0$, and $P(b \mid a) = P(b)$.

Relevant marketing mix variables for the purchase incidence stage include the category attractiveness (w'_a), household inventory, seasonality, category price, environmental factors such as trend and inflation. Also, an important variable to model for multiple unit purchases is a dummy to indicate whether the current purchase incidence corresponds to a purchase decision of the first unit or of the second or further unit within the given shopping trip. Examples of marketing variables for the brandsize choice stage are price and promotions, which are characteristic of the products, and advertising exposure, coupon, previous purchase history, and brandsize preference, which are characteristic of the households.

The estimation can be carried out by the maximum likelihood method. Although a full information joint estimation is possible, a convenient and consistent sequential estimation permits a use of the standard multinomial logit routine with some loss of efficiency. (McFadden 1981, Ben-Akiva & Lerman 1985)

2.2 The Utility Residual Method (URM)

URM introduced in Part II generalizes the standard linear-in-parameter utility function in the multinomial logit model by additive nonparametric utility function of each covariate as

(2) $$P_j = \frac{e^{v_j}}{\sum_k e^{v_k}} \qquad \text{where} \qquad v_j = \sum_p \phi_p(x_{jp})$$

where x_{jp} is a p-th covariate for alternative j and $\phi_p(\cdot)$ is a one dimensional nonparametric function. We will utilize the method in a conservative manner as a diagnostic tool and choose better parametric transformations of the covariates for the utility function, if necessary, as a result of subjective judgements on $\phi_p(\cdot)$. Because URM does not automatically capture the interaction among covariates, these aspects still need to be manually examined with care.

3. RED DRINK SINGLE SOURCE DATABASE

A product category studied is so called Red Drinks which includes cranberry cocktail and any blends of cranberry such as cranberry apple and cranberry raspberry. In this category, a share model (i.e., brand choice model) describes only part of the market phenomena because of many switchings between other drink categories such as fruit juices and soft drinks. A sales model describing the category expansion and contraction is crucial to describe the whole picture. The database is a single source data from Grand Junction, Colorado supplied by IRI, and contains panel records including shopping trips and store records over two years (10/12/87~10/08/89), and TV advertising exposure data monitored by TV meters for the last one year.

15 highest share brandsizes which constitute 71% of the category purchases are extracted for the study. The products are listed in Table 2 with their shares and average prices of the purchases. Ocean Spray is the only national brand appearing in the top 15 and the other two brands, FD and JL are both private (store) labels. The highest ranking occupied by any other national brand is only 30th. We also note that Ocean Spray is the only TV advertiser for the category in the market although local newspaper features may appear for other brands.

194 households who meet IRI continuity criteria as sample[11], are continuously monitored over the entire two years for their shopping trips and purchases, while their TV data contains date

[11] Households who drop out or join in the middle of the sample period or who have overly long gaps between shopping trips are excluded. However, it does not imply that the households are regular purchasers of the category products. In fact, there are 54 households with only a single category purchase over the last two years, which made our calibration quite challenging.

and daypart of each Ocean Spray media exposure. Numbers of various events in the database are summarized in Table 3.

Table 2: Shares and average prices of each brandsize in the Red Drink Database

brandsize	share (%)	average price (¢/oz)
FD cranberry 32oz	6.25	4.34
FD cranberry 48oz	9.26	4.30
JL cranberry 48oz	3.75	3.44
JL cranapple 48oz	3.09	3.44
JL cranraspberry 48oz	3.09	3.42
OS cranapple 48oz	8.75	4.81
OS cranapple 64oz	4.41	4.57
OS cranapple 128oz	3.75	4.35
OS cranberry low-cal 48oz	6.76	4.82
OS cranberry 32oz	11.32	5.46
OS cranberry 48oz	12.35	4.79
OS cranberry 64oz	10.37	4.57
OS cranberry 128oz	5.59	4.31
OS cranraspberry 48oz	6.47	4.78
OS cranraspberry 64oz	4.78	4.59

Table 3: Statistics of various events in the sample database

period	10/13/86~10/11/87	10/12/87~10/09/88	10/10/88~10/08/89
weeks	52	52	52
purchases	371	410	579
trips	n.a.	23,637	23,635
purchase opportunities [12]	n.a.	24,073	24,239
ad exposures	n.a.	n.a.	7,474
store data	n.a.	yes	yes

*** n.a. = not available**

[12] Refer to section 5.1 of the category model.

In the next two sections, we will construct a brand choice and category purchase model which will be calibrated and validated. Due to the limited time frame on certain data, the procedure will be as follows. Household purchase data without trip or store information for the year preceding the sample period is used as pre-calibration for initializing certain variables (e.g. loyalty, household inventory). Next, to allow for the maximal accuracy, the entire two years is used for selecting covariates and obtaining their transformations. Then, the model is calibrated based on the first 72 weeks of the two years (10/12/87~2/26/89) with 628 purchases and 32,573 trips and tested on the holdout sample during the remaining 32 weeks (2/27/89~10/02/89) with 361 purchases and 14,699 trips. In both cases, a null model of equal probabilities is used for a measure of fit, ρ^2 and $\bar{\rho}^2$. All computations are done on a Dell 486 PC.

4. BRAND CHOICE MODEL

4.1 Brand Choice Model Specification

Our main objective here is to build a parsimonious marketing mix model with good predictive fit. The following variables are selected after extensive data analysis.

Brandsize loyalty (between 0 and 1)
Feature (0, 1, 2, or 3)
Display (0 or 1)
Price (cents/oz)
Adstock (non-negative)

Brandsize loyalty for the j-th alternative defined as

(3) $$loyalty_j(t+1) = \lambda_1 \cdot loyalty_j(t) + (1-\lambda_1) \cdot d_j(t)$$

where $d_j(t)$ is 1 if alternative j is bought at t-th purchase occasion, 0 otherwise,

is similar to the brand and size loyalty in Guadagni & Little (1983) But the value is associated for each alternative (brand and size pair) instead of brand and size separately. Use of the brandsize loyalty imposes less constraint than using brand and size loyalties, and tends to

produce better fit with parsimony.[13] Due to the fact that the loyalty takes into account cross-sectional and time series variation with the same decay constant, some researchers propose to decompose into two elements: one for stationary household idiosyncrasies and the other for their dynamic change. (Fader & Lattin 1990) Since at least in some cases, such a decomposition does not have much impact on prediction or estimation of other parameters (Guadagni & Little 1987), and since our database does not contain enough purchases to allow accurate estimation of stationary preference for many panelists, we will stick with the single measure. The decay constant, λ_1, is estimated to be 0.774 by the Taylor series method (Fader, Lattin, & Little 1990) which maximizes the likelihood function.

Feature is a promotional product feature in local newspapers and store circulars, and takes a value of either 0, 1, 2, or 3 depending on its effectiveness. (IRI classifies them as C, B, and A-ad) Display is a binary in-store display indicator. Price is defined as price paid at a cashier divided by the volume in cents per ounce. It includes both promotional price-cuts and coupons. Decomposing the price into regular unit price and unit price-cut did not improve the fit, and

[13] Separating brandsize loyalty into brand and size loyalty or brand, size, and flavor loyalty resulted in worse fit as shown below.

specification	1	2	3
$\rho 2$	**0.4755**	**0.4597**	**0.3669**
ρ^2-adjusted	**0.4684**	**0.4519**	**0.3594**
brandsize loyalty	5.474 (29.85)	—	—
size loyalty	—	2.956 (17.91)	3.282 (20.26)
brand loyalty	—	2.087 (9.13)	2.523 (10.72)
flavor loyalty	—	3.144 (18.69)	—
feature	0.372 (3.29)	0.360 (3.17)	0.367 (3.35)
display	1.071 (6.06)	1.058 (5.89)	0.841 (4.94)
price	-0.985 (-4.77)	-1.163 (-5.44)	-0.994 (-4.93)
adstock48	0.116 (2.90)	0.112 (2.70)	0.101 (2.56)

* N = 989

simply separates the price effect into two components. Coupon usage is very light and less than 2.5% of the purchases (31 purchases) are accompanied by coupons. In addition, because coupon availability is not known, we presume their effect to be similar to discounting and let it be absorbed in the price variable. None of the two way interactions among feature, display and price (or even price-cut) are significant.

The television commercial broadcasts (three types) for Ocean Spray are primarily brand image oriented and do not strongly differentiate among flavors. Thus, we assume the same advertising exposure level for all Ocean Spray brandsizes regardless of their flavor and size. There was no private label TV ad. The adstock variable for each panelist was constructed as a sum of all previous exposures encountered before a particular purchase occasion in question, adjusted for memory recall (forgetting) by daily decay. The variable is constructed based on behavioral and field studies of advertising (Little & Lodish 1969, Lodish 1971, Clarke 1976, Craig, Sternthal, & Leavitt 1976), and captures a direct carry-over of advertising effect in continuous time frame unlike Tellis (1988), Kanetkar, Weinberg, & Weiss (1989), and Pedrick & Zufryden (1990). The estimated decay rate per day by the Taylor series method produced 0.914, which corresponds to 53% retention after a week. Lodish (1971) found a slightly slower decay of about 70% retention after one week and 30% after 4 weeks for magazine ads from several empirical studies. We have created size and flavor specific adstock by multiplying 0/1 indicator dummy for appropriate brandsizes.

Motivated by the study of Tellis (1988), interaction terms between brandsize loyalty and a linear and quadratic size-specific adstock were introduced in the model. Neither of them was significant as is the case in his article. Similarly, we have introduced price and adstock interaction to examine whether advertising affects price sensitivity. Kanetkar et. al. (1989) find a significant negative coefficient and thereby conclude increased price sensitivity in their dry dog food and aluminum foil product category. The effect was not observed in our study which found an insignificant positive coefficient (t=0.23).

Table 4 is the result of models with various adstock variables calibrated on the whole two years for the purpose of variable selection. In general, adstock coefficients are much weaker than those of other marketing mix variables --- a fact observed in other scanner studies (Tellis 1988). Advertised flavors receive a slightly stronger effect than non-advertised flavors, but the difference is almost negligible. Among different sizes of Ocean Spray brands, 48oz --- the primary size for the category --- indicates a significant positive, while relatively insignificant negative is observed for 32oz. Adding adstock quadratic terms as in Tellis to account for the

saturation phenomenon did not change the general relationship among the adstock coefficients for different sizes.

Table 4: Result of MNL models with various adstock variables

specification	1	2	3
$\rho 2$	**0.4755**	**0.4761**	**0.4745**
$\bar{\rho}^2$	**0.4684**	**0.4679**	**0.4671**
adstock32	-----	-0.102 (-0.79)	-----
adstock48	0.116 (2.90)	0.139 (1.97)	-----
adstock64	-----	-0.017 (-0.21)	-----
adstock128	-----	0.096 (1.12)	-----
adstock flavor	-----	-----	0.096 (1.34)
adstock non-flavor	-----	-----	0.061 (0.95)
brandsize loyalty	5.474 (29.85)	5.477 (29.74)	5.446 (29.66)
feature	0.372 (3.29)	0.361 (3.19)	0.385 (3.40)
display	1.071 (6.06)	1.084 (6.10)	1.049 (5.95)
price	-0.985 (-4.77)	-1.000 (-4.82)	-0.994 (-4.80)

*** N = 989**

Since one desirable model characteristic is parsimony, only significant and semi-significant variables are retained throughout the study. Consequently, the brand choice model includes five variables: brandsize loyalty, feature, display, price, and adstock for 48oz, in addition to 14 alternative specific constants which are kept to avoid bias in the coefficients.

4.2 Analysis by URM

Finally, URM is applied to obtain better behavioral insight and to examine whether any variable transformation is necessary. The loyalty plot shown in Figure 1 is very similar to those observed in the two previous applications in Part II. The curve suggests that there exist three segments, non-buyers and loyal buyers of a brandsize with a steep utility slope and switchers with a very flat region. To model the curve in a parsimonious way, transformation by a cubic function with the inflection point at 0.45 is applied. The parametric function approximates the URM transformation very well when scaled by the estimated coefficient in Figure 2.

The price plot of Figure 3 resembles the curve presented by Gurumurthy & Little (1989) in that it has a relatively flat portion around the most frequently found prices. As seen in Table 2, the price range of most brandsizes is between 4 and 5 cents/oz, in which consumers' utility is insensitive to price movement. If price is lower than 4 cents/oz because the product is either discounted or JL private label, utility increases quickly, while price higher than 5 cents/oz results in steeper decreasing utility. Trying to fit a cubic function as in the loyalty case, however, led to lower fit and thus, we leave it linearly specified.

The adstock plot in Figure 4 suggests a saturation effect of advertising, which is observed in many studies. (Little 1979) Logarithmic transformation, log(adstock+1), shown in Figure 5 is employed to account for the diminishing return.

The model with loyalty and adstock transformations results in improving fit of 0.0034 in ρ^2. In the following, we will compare both the linear and parametrically transformed models in the calibration and holdout sample.

4.3 Calibration and Validation

Table 5 is the result of the linear and transformed model on the calibration sample. The feature variable which is significant at t=3.3 using all 104 weeks becomes insignificant. We suspect that this is due to sample variation. The transformed model has a slightly higher ρ^2, but lower mean probability and percent of correct choices. Once again, this implies robustness

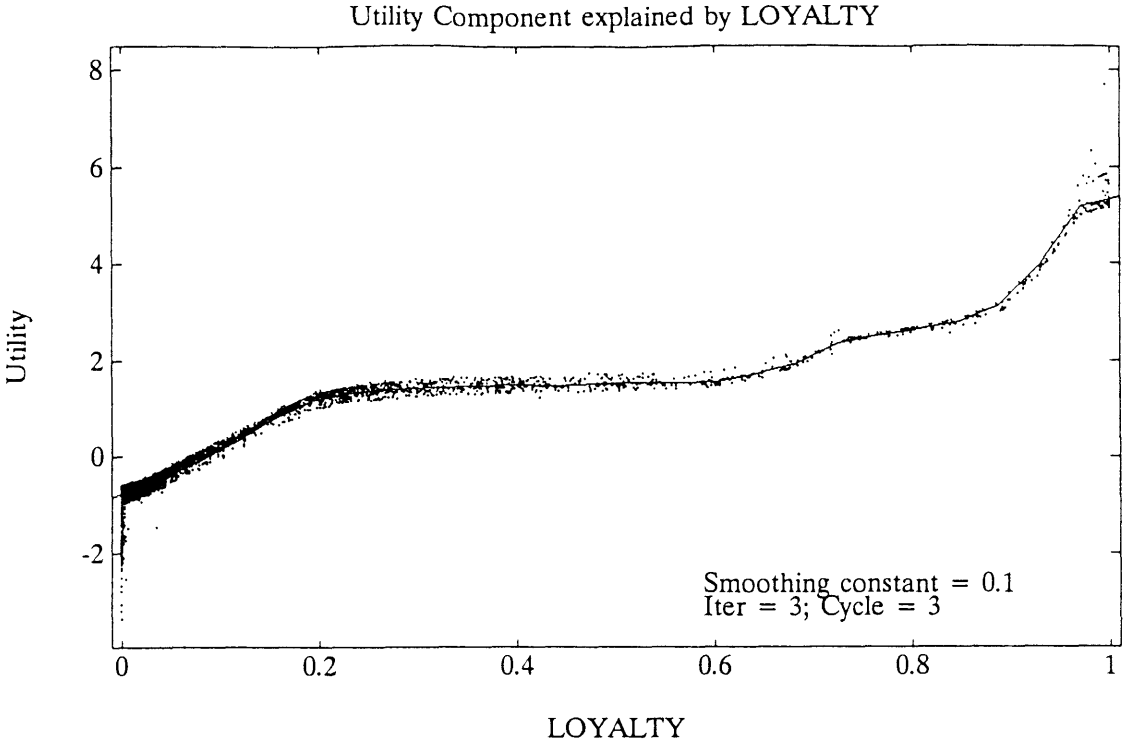


Figure 1: Loyalty utility transformation by URM

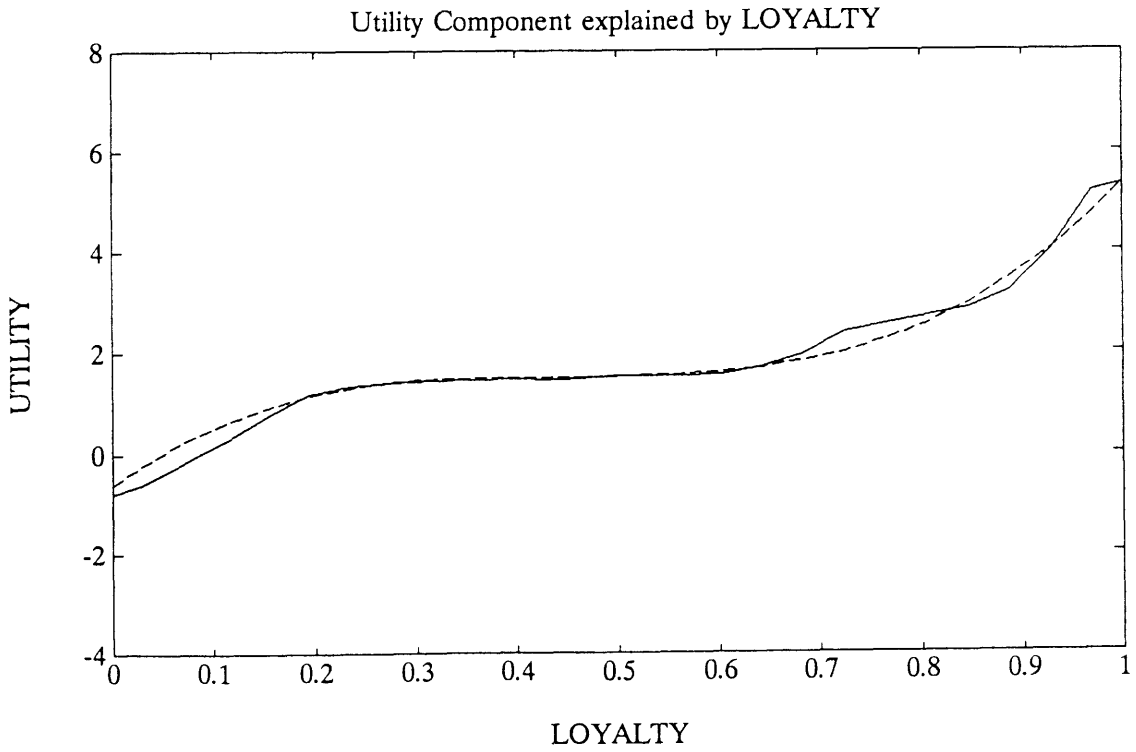


Figure 2: URM vs. parametric transformation of loyalty variable

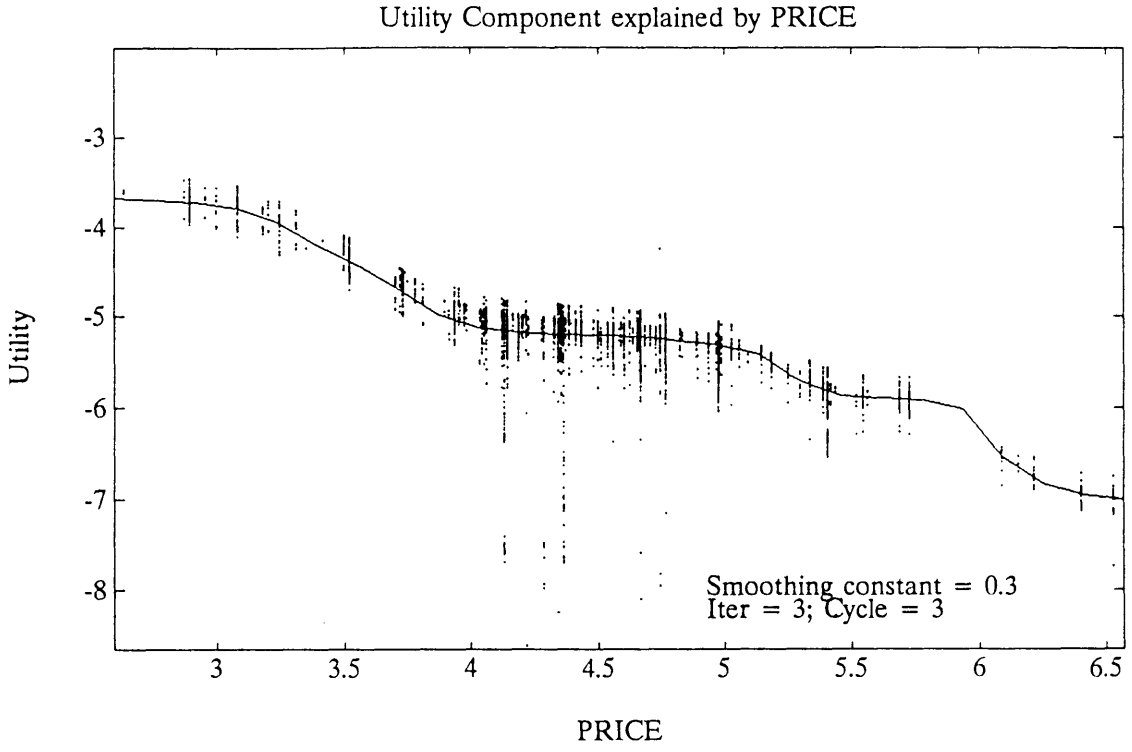


Figure 3: Price utility transformation by URM

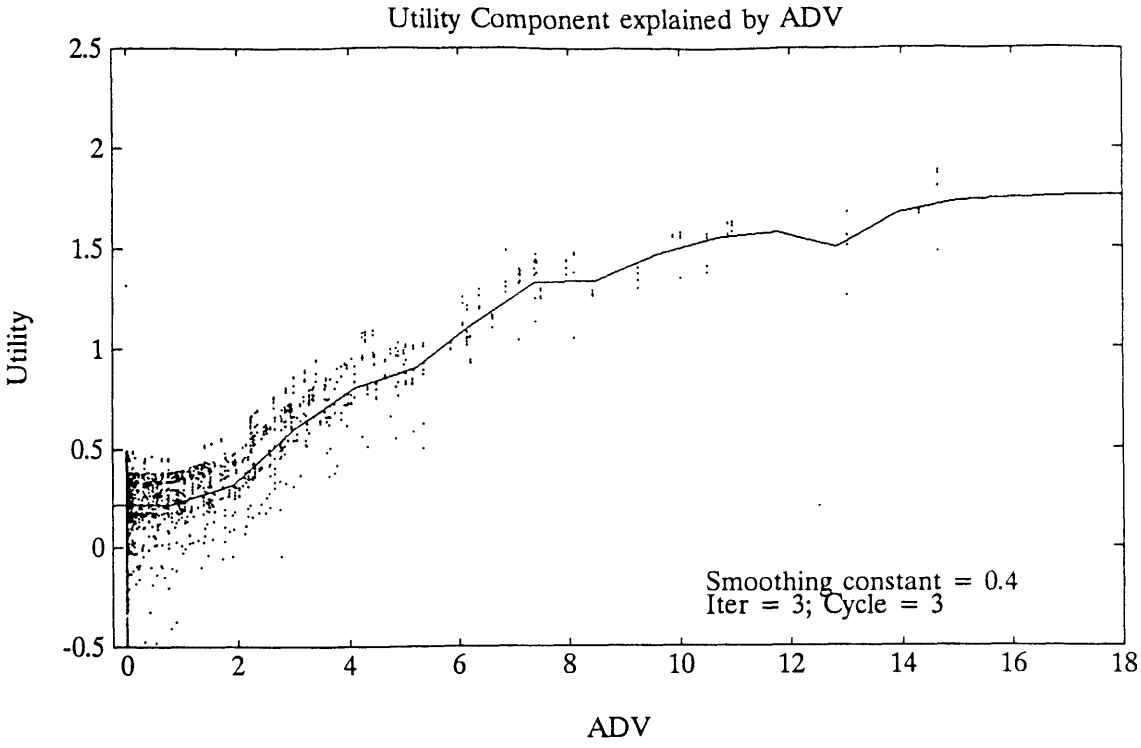


Figure 4: Adstock utility transformation by URM

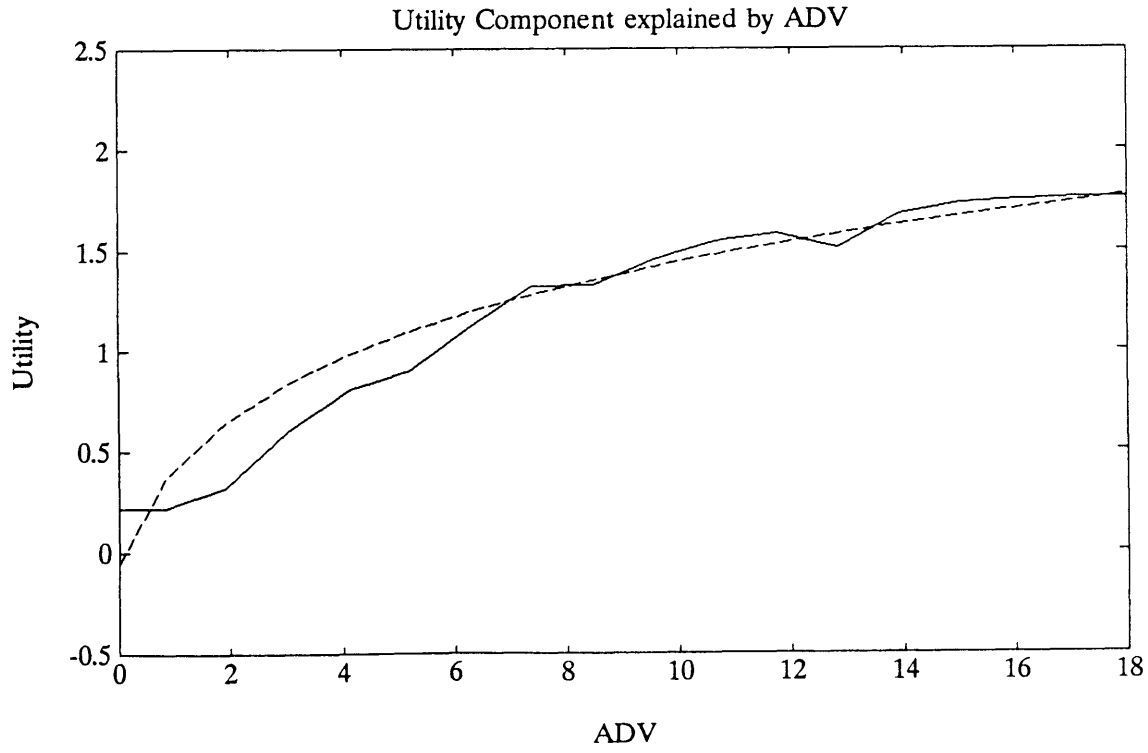


Figure 5: URM vs. parametric transformation of adstock variable

of the linear model in prediction even when the utility function is incorrectly specified as demonstrated by the simulation study of URM in Part II.

Table 5: Calibration result of brand choice model

specification	Linear	Transformed
$\rho 2$	**0.4851**	**0.4895**
$\bar{\rho}^2$	**0.4739**	**0.4783**
prob. of correct choice	0.425	0.424
percentage of correct choice	54.30	53.66
brandsize loyalty	5.582 (23.17)	-----
adstock48	0.124 (2.75)	-----
(loyalty - 0.45)^3	-----	28.28 (20.93)
log (adstock48 + 1)	-----	0.485 (2.28)
feature	0.164 (0.84)	0.140 (0.71)
display	1.021 (4.17)	0.960 (3.92)
price	-1.063 (-4.06)	-1.046 (-4.04)

*** N = 628**

Because the loyalty variable involves the carry-over based on previous purchases, for projection in the holdout sample, a purchase sequence predicted by the model must be used as purchase history. As in Guadagni & Little (1983, 1987) and Gupta (1988), this is done by repeated runs of Monte Carlo simulation to stabilized the outcome. Here, we report the result based on 50 runs. Table 6 confirms the validity of the two parsimonious models by comparing fitting criteria between the calibration and holdout sample. R^2 is sample correlation of 4-week shares between the actual and predicted for all brandsizes. Time series share plots of the total

Ocean Spray and the major brandsize --- Ocean Spray Cranberry Cocktail 48oz --- by 4-week period are shown in Figures 6 and 7 respectively for the linear and transformed model. A visual inspection suggests a better fit by the linear model in the holdout as well.

Table 6: Goodness-of-fit for linear and transformed model in calibration and holdout sample

model		prob. of correct choice	ave. loglikelihood	R^2
Transformed	*calibration*	0.424	-1.382	0.671
	holdout	0.402	-1.482	0.616
Linear	*calibration*	0.425	-1.394	0.682
	holdout	0.414	-1.488	0.651

Due to the mixed result, we adopt the standard and parsimonious linear model for brand choice, and the category attractiveness variable used in the category purchase model will be computed accordingly.

5. CATEGORY PURCHASE MODEL

5.1 Category Purchase Model Specification

Because relatively few disaggregate category purchase studies have been done in the past (Guadagni & Little 1987, Gupta 1988) unlike brand choice, we propose the following variables for possible inclusion in the model and proceed cautiously using a priori hypotheses and empirical testing.

buy-later dummy (0/1)
first purchase opportunity (0/1)
category attractiveness (unitless)
category price (cents/oz)
adstock (non-negative)

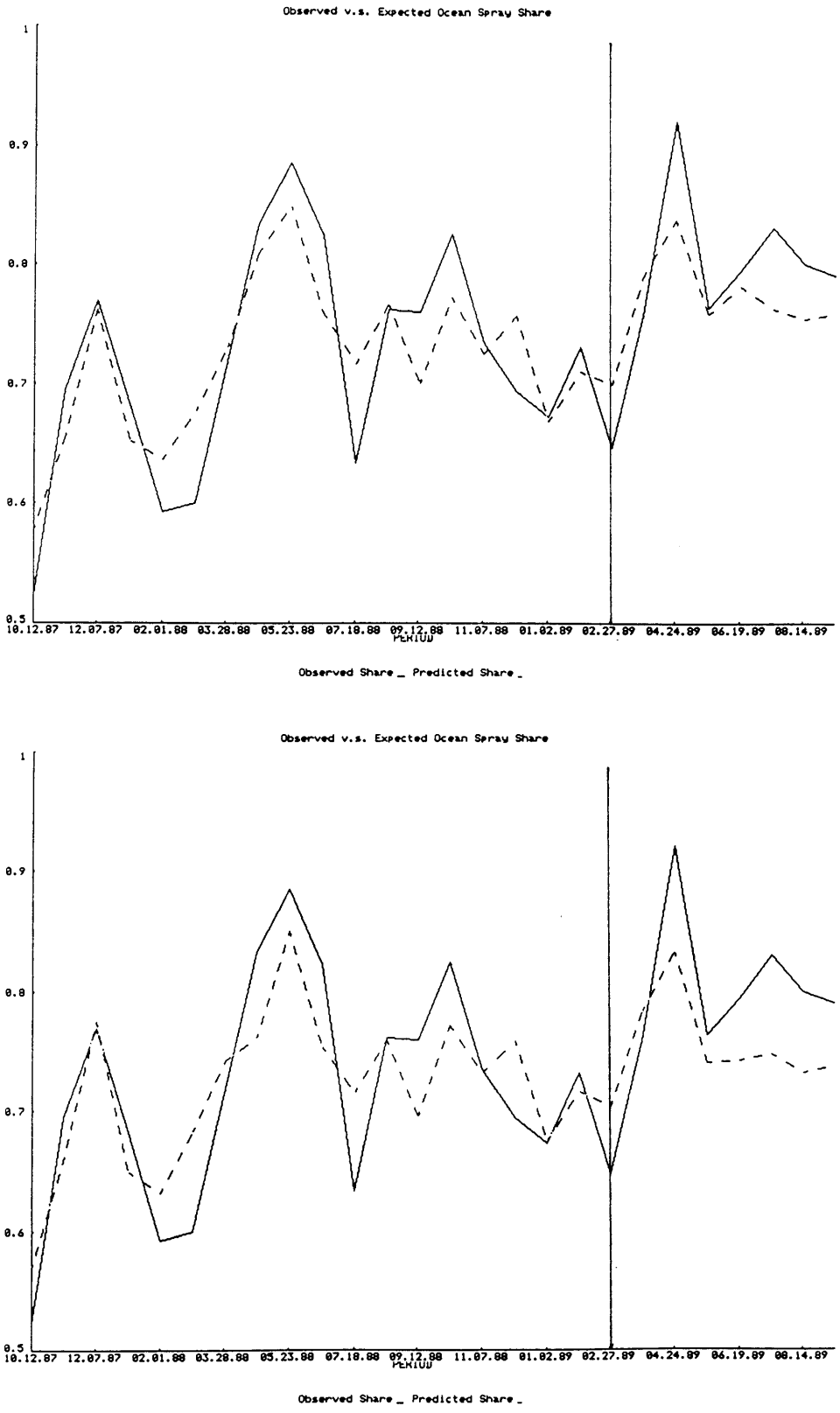


Figure 6: Time series plots of the total Ocean Spray share by linear and transformed brand choice model (top: linear, bottom: transformed)

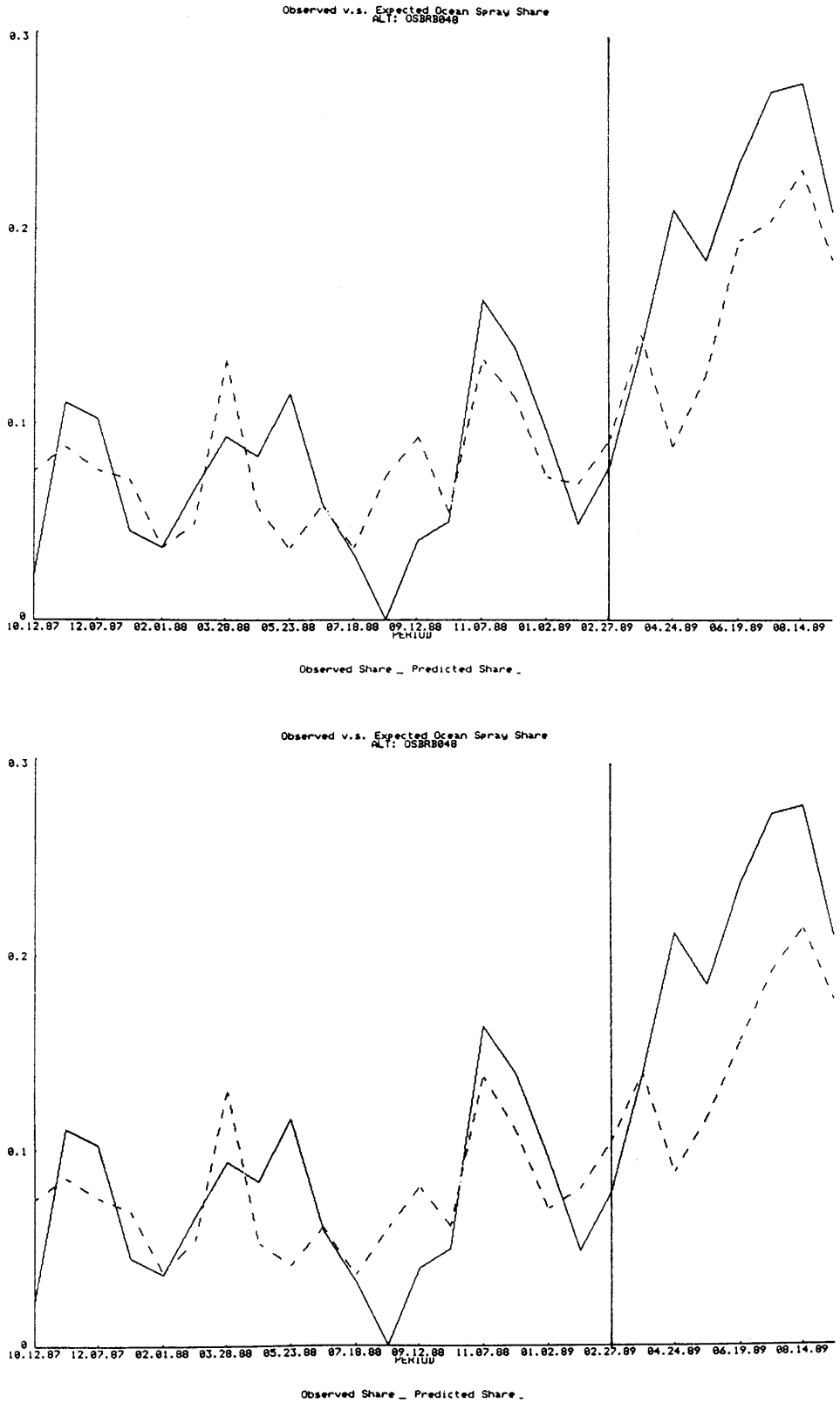


Figure 7: Time series plots of Ocean Spray cranberry 48oz share by linear and transformed brand choice model (top: linear, bottom: transformed)

category attractiveness on multiple unit purchases (unitless)
standardized spending on a shopping trip (unitless)
household buying rate (oz/week)
household consumption rate (oz/week)
household inventory (weeks of supply)
seasonal index (unitless)

Values for these variables are specified for each of the two alternatives --- buy-now and buy-later, and for each purchase opportunity which is defined as in Guadagni & Little (1987). That means the number of purchase opportunities is the number of trips plus the total number of units purchased to account for multiple unit purchases (45 out of 989 purchases).

Buy-later dummy is an alternative specific constant for the binary logit. ***First purchase opportunity dummy*** indicates whether the particular purchase opportunity is for the first unit or the second and further, and captures the likelihood of multiple unit purchases not explained by other variables.

Category attractiveness is the expected maximum utility of all brandsizes faced at a shopping trip (purchase opportunity) by a household. It is an output from the subtree in the nested logit formulation described in section 2.1 and also referred to as the "inclusive value". Its value for buy-later is an 8-week moving average of buy-now category attractiveness as in Guadagni & Little. We have also created category price and adstock separately to examine their unique roles not taken into account by the category attractiveness.

Category price for buy-now is an average of brandsize prices[14] at a particular purchase opportunity, and its 8-week moving average is used for buy-later counterpart. When introduced in the full model, the category price was not significant (t=-0.97) and lowered the t-value of the category attractiveness by 1.5. In addition, because correlation between the two variables are quite high (-0.63), we decided not to include the category price in the model.

Adstock is the one used in our brand choice model, except that the decay is now updated for time of each trip. When introduced, the adstock was insignificant at t=-0.60 and had moderately high correlation with the category attractiveness (0.32). Hence, it was also dropped.

---

[14] We did not take the average weighted by a household preference measure such as loyalty done by Gupta (1988) because we wished to avoid collinearity with the category attractiveness.

Category attractiveness on multiple unit purchases is created to examine the effect of price and promotions on multiple unit purchases. It is set equal to category attractiveness only on purchase opportunities for the second or further purchases and to zero otherwise. Equivalently, it is an interaction term between category attractiveness and one minus the first purchase opportunity dummy. The variable captures the tendency (if any) to buy an extra unit (given that at least one unit has already been bought) as a result of the difference between the current utility of the product --- determined by the current price, promotions, adstock, and loyalty--and its 8-week moving average. In other words, a significant positive sign suggests a stockup of a temporally promoted and/or advertised product.

Standardized spending on a shopping trip is defined as a dollar amount spent on a shopping trip divided by a household specific average spending of all shopping trips.[15] The standardization takes care of heterogeneity between heavy and light consumers characterized by, for example, family size.

A household purchase rate is decomposed into two components: cross-section across households (heterogeneity) and time series within a household (longitudinal effect). The ***household buying rate*** --- defined as the total volume of category purchase by the panelist divided by the number of weeks over the entire three years --- accounts for the former. The latter dynamics is captured by a ***household consumption rate*** with the exponential smoothing of the past buying rates as,

(4) smoothed consumption rate (at current category purchase) =
$\lambda_c \cdot$ smoothed consumption rate (at the previous category purchase)
$+ (1-\lambda_c) \cdot$ (volume of the previous purchase / smoothed interpurchase time)

where λ_c is estimated to be 0.490. The starting consumption rate is set to be the household buying rate above. The smoothed interpurchase time in (4) is computed to account for its time-varying effect as,

(5) smoothed interpurchase time (at current trip) =
$\lambda_t \cdot$ smoothed interpurchase time(at previous category purchase)
$+ (1 - \lambda_t) \cdot$ (most recent interpurchase time) .

[15] This variable was suggested by D. Honnold of IRI.

Equation (4) and (5) represent the same smoothing scheme used for the loyalty variable in the brand choice model, and is equivalent to averaging with geometric weighting to adapt more to recent events. That is, a smoothed variable, y, is expressed as the weighted mean of past covariate variable, z, as

$$(6) \qquad y(t+1) = \lambda\, y(t) + (1-\lambda)\, z(t) = \frac{\sum_{k=0}^{\infty} \lambda^k z(t-k)}{\sum_{k=0}^{\infty} \lambda^k}$$

The decay constant, λ_t, for the interpurchase time is estimated to be 0.73 by the Taylor series method. The average household interpurchase time over the three years, (i.e., [no. of weeks] / [no. of category purchases]), is used to start the variable. This average is also used for panelists with a censoring problem due to few purchases (less than 3).

The value of the consumption rate for buy-later is set to be the household buying rate, which is a stationary measure of a consumption rate.

The ***household inventory*** is defined in unit of weeks of supply to account for household consumption heterogeneity as in Guadagni & Little. Then, the inventory at the r-th purchase opportunity is,

$$(7) \qquad \text{inventory}(r) = \text{inventory}(r-1) - \frac{[\text{date}(r) - \text{date}(r-1)]}{7} + \frac{\text{volume purchased } (r-1)}{\text{usage rate } (r)}$$

We have two candidates for the usage rate in (7) from the previous definitions. One is the *household buying rate* which is stationary and the other is the *household consumption rate* which is dynamic. The stationary one resulted in much better fit than the latter by 0.0095 in ρ^2. The reason seems to be the strong correlation between the dynamic inventory and the household consumption rate which is also time-varying.

We do not clip the values of the household inventory at either the upper or lower bound unlike Guadagni & Little (1987) and Gupta (1988). Therefore, if the category purchase is not made in

a long time, the inventory could be driven towards a large negative value. At first, this may not seem reasonable if only physical product inventory is concerned. However, the definition of this inventory perhaps captures "mental state of inventory", i.e., a desire to consume the category product because one hasn't used it for a long time. Taking into account this behavioral phenomenon is potentially important in diverse categories such as Red Drinks where consumers can readily seek variety and substitutes by other drink categories. This is in contrast to a coffee category studied by the aforementioned articles since coffee is very much like necessity for many addicted coffee drinkers so that the physical inventory overshadows the mental one. The data also supports this formulation with much better fit (improvement of 0.0280 in ρ^2) than when the clipped inventory is used. The base level of the inventory for a household is shifted in such a way that the inventory becomes non-negative by adding the most negative value to all observations. The household inventory for buy-later is 0 as in Guadagni & Little.

Finally, a plot of category purchase volume by week over the two years indicates influences from seasonality and certain holidays, but not from steady trend --- either expansion or contraction. Thus, we construct a weekly ***seasonal index*** as follows to avoid any confounding with marketing mix activities. First, the weekly number of category purchases is predicted by the category model with all covariates except seasonal index. Then, these predictions are subtracted from the observed number of purchases to form residuals on weekly basis. The seasonal index is an average of the two residuals which correspond to the same week of the two year.[16] The resulting seasonal index which is normalized to have a mean of 0 is shown in Figure 8. Difference between a peak (Nov.~Apr.) and off-peak (May~Oct.) season by an IRI definition can be observed. A drop right after the Christmas --- a phenomenon clearly found in the data --- is also reflected in the plot. The index for buy-later is set to zero.

To summarize, the category purchase model has nine variables: buy later dummy, first purchase opportunity, category attractiveness, standardized spending on a shopping trip, category attractiveness on multiple unit purchases, household buying rate, consumption rate, inventory, and seasonal index.

---

[16] It is not possible to construct a seasonal index without using the holdout data because minimum duration of 2 years is necessary. Due to this construction, prediction performance in the holdout is expected to be somewhat inflated.

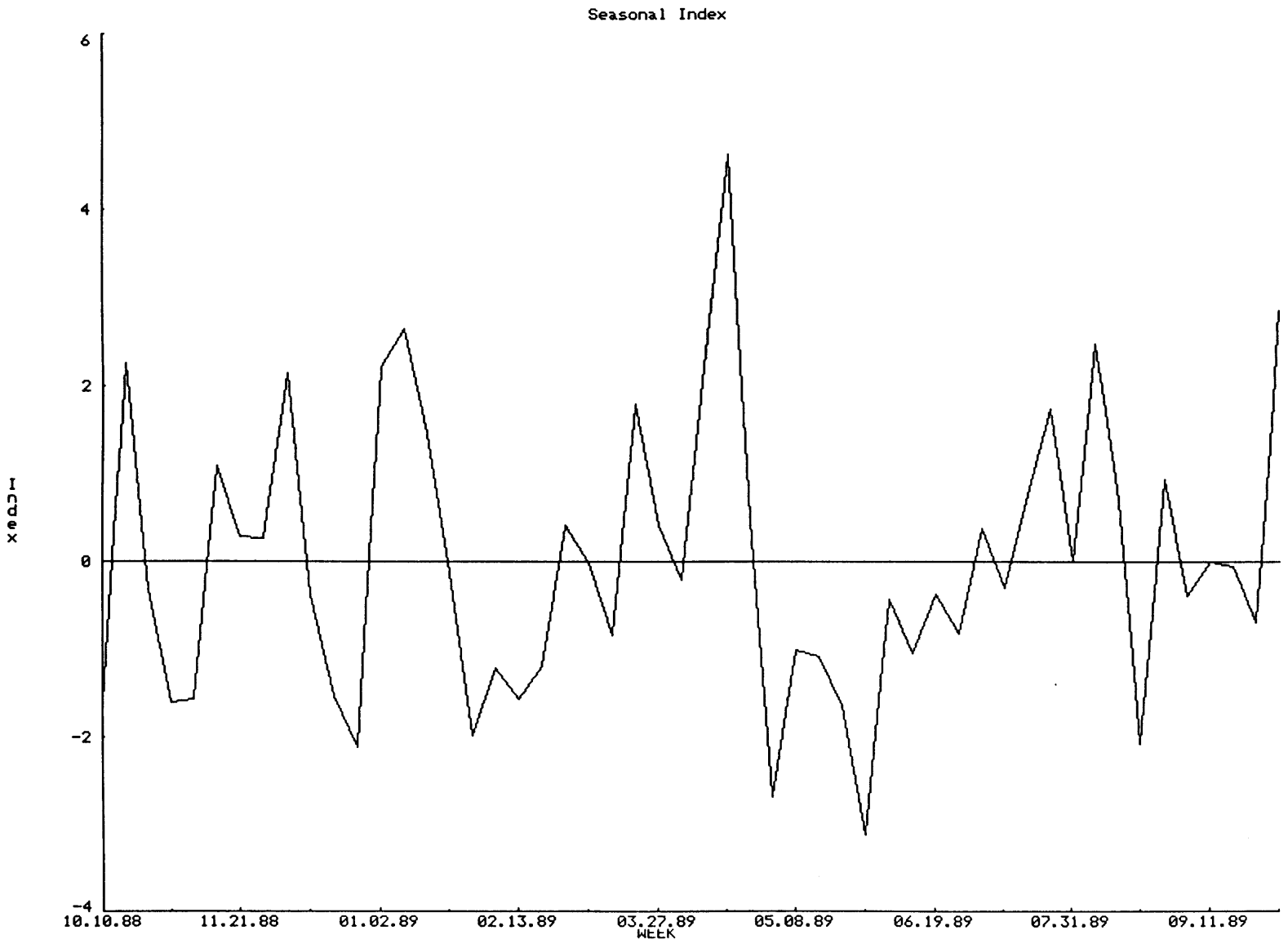


Figure 8: Seasonal index plot

5.2 Analysis by URM

As shown in Table 3, the total number of category purchases is 1,040 among 48,312 purchase opportunities over the two years. This corresponds to only 2% of buy-now choices for the observations, and much computational effort is expensed for estimating the buy-later dummy in the binary logit. Therefore, we introduce choice-based sampling in the category purchase model estimation to reduce computation time while still maintaining relevant information. The choice-based sampling refers to a sampling scheme based on the chosen alternative. In our case, the sample consists of all purchase opportunities with buy-now observations (1,040 purchase opportunities) as well as random sample within the buy-later observations (3,771 purchase opportunities). The fraction of buy-now observations constitute approximately 20% of the reduced sample, which is now only one tenth in size of the original data. Details about the choice based sampling appear in Ben-Akiva & Lerman (1985). See also McFadden (1981) who proves that, for a multinomial logit model with a full set of alternative specific constants, the maximum likelihood procedure on the choice-based sample yields the consistent estimates of all parameters except the constants. Consequently, we gain a computational advantage for some minor loss in estimation efficiency but not consistency.

The smaller sample size makes the analysis by URM on a 486 PC much more manageable with computational time of about 20 minutes. Figure 9 is the initial empirical probability plot before the actual URM algorithm starts (cycle = 0). There are two influential points in the symmetrical locations --- one at (y=0, w≈9) and the other at (y=1, w≈-9), which are due to nature of the binary logit model.[17] As a result, the kernel regression is strongly distorted in these regions. The two points turn out to be purchase opportunity index 18,359 whose household consumption rate is 322.77 oz/week corresponding to the upper bound of observed values (see Figure 11) but no category purchase is made. Hence, these points are removed from the sample. This illustrates an advantage of graphical diagnostics made possible by URM in logit models.

The resulting nonparametric utility transformations by URM are presented in Figures 10 through 13. First, notice that there are many more outliers than in the brand choice. We must focus our attention to the portion of the curve where many observations fall and disregard the regions of the outliers.

The inventory plot shows a discontinuous utility curve, and it seems to suggest that there exist two rather distinct segments. To make sure that the phenomenon is not a spurious effect caused

[17] In the case of binary logit, $w_1 = v_1 - v_2$, according to the definition of w_1 in (4a) of Part II.

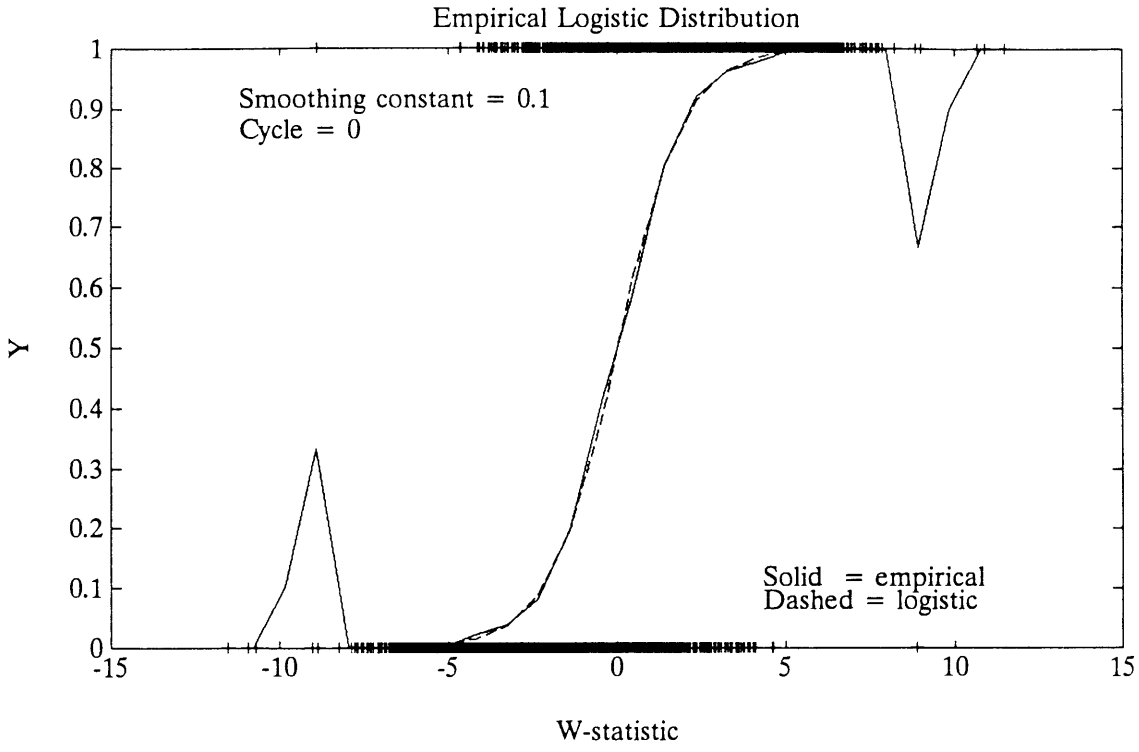


Figure 9: Empirical v.s. theoretical probability plot by URM

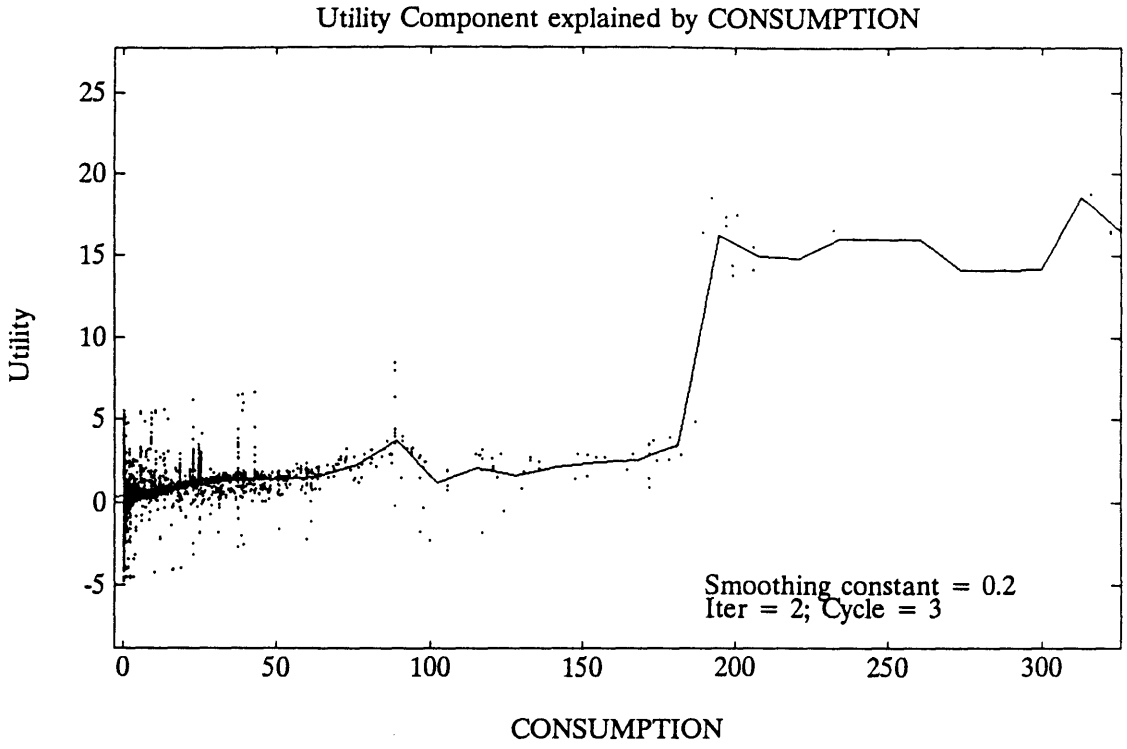


Figure 10: Household consumption utility transformation of the whole choice-based sample by URM (N=4810)

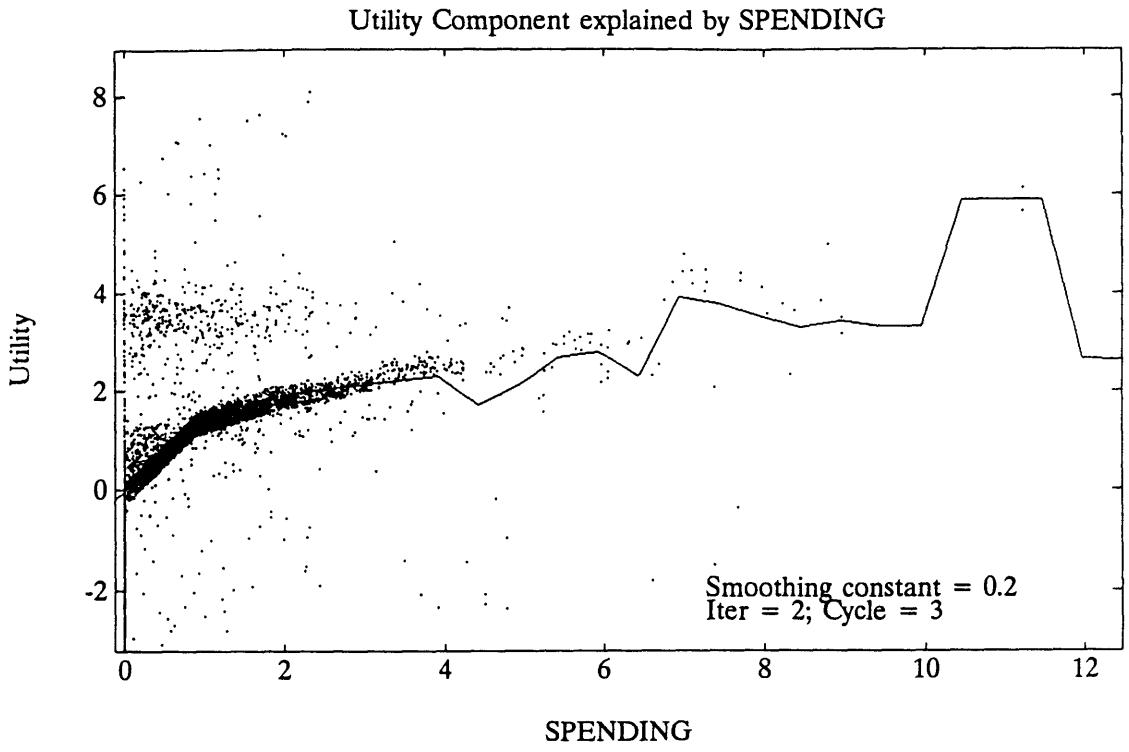


Figure 11: Standardized spending utility transformation of the whole choice-based sample by URM (N=4810)

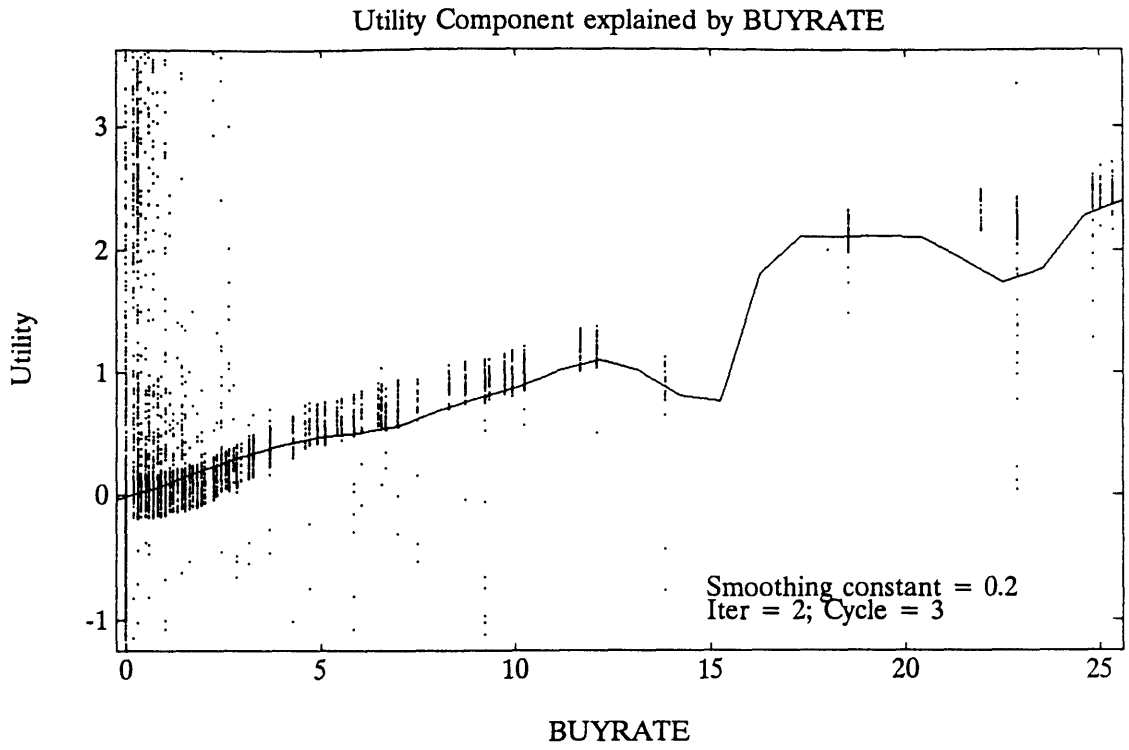


Figure 12: Household buying rate utility transformation of the whole choice-based sample by URM (N=4810)

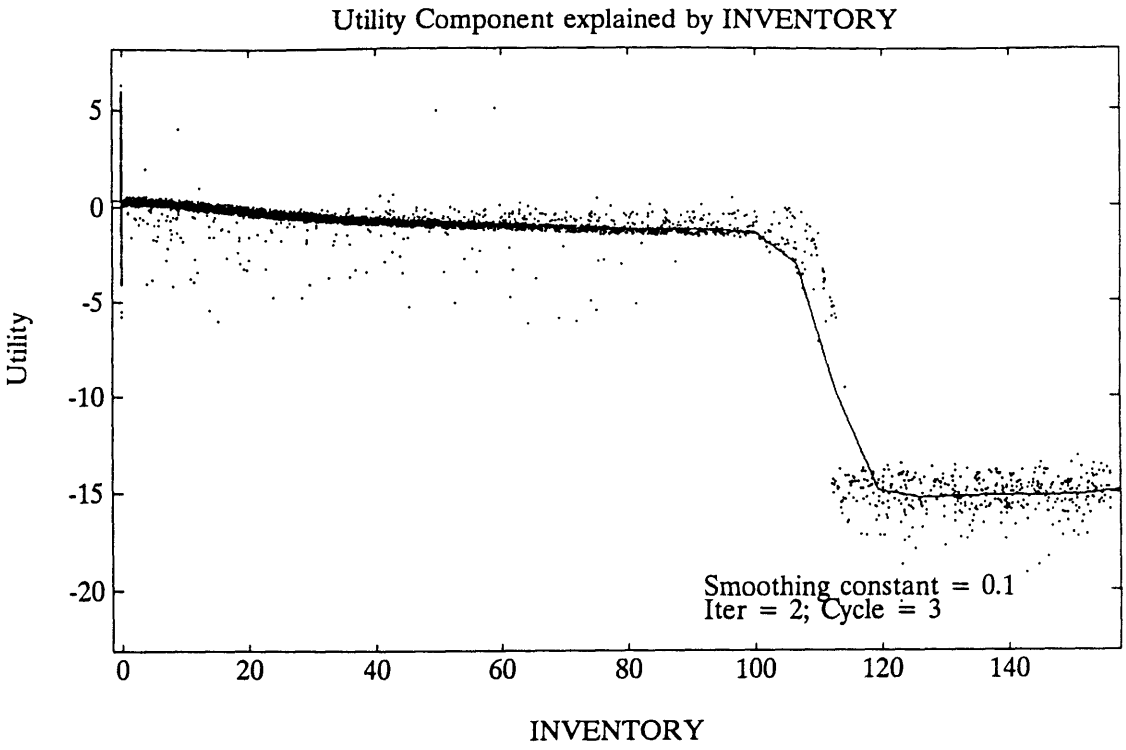


Figure 13: Household inventory utility transformation of the whole choice-based sample by URM (N=4810)

by the other transformations, only the inventory is allowed to be nonparametric while other variables are held as linearly specified. Still the same discontinuity was observed, which is shown in Figure 14. The lower right segment with inventory greater than about 110 weeks of supply corresponds to infrequent buyers. Thus, we decided to drop households who made less than three purchases during the sample period of two years and re-estimated the category model.

Now, the number of observations is reduced to 2,223, of those, 770 are buy-now and the remaining 1,453 are buy-later. The resulting URM transformations are shown in Figures 15 through 18.[18] The inventory no longer exhibits the two segments. Focusing the attention to areas where points are concentrated and taking into account the boundary effect of kernel regression which tends to flatten the curve (Part II), linear specification seems appropriate for all but the standardized spending variable. Because the spending exhibits a diminishing return in the dense region between 0.25 and 2.5, a logarithmic transformation, log(spend+1), was tried, which resulted in an improvement of 0.017 in ρ^2. Figure 19 shows the parametric function along with the URM curve after rescaling by the estimated coefficients. For other variables, quadratic terms were added to capture the possible curvatures, but none of them were significant. Therefore, household inventory, consumption rate, and buying rate are left to be linear, and we evaluate both the linear and logarithmic standardized spending models in the calibration and validation.

[18] Because points are sparse at the higher end of the scale for each variable as seen in Figures 10 to 13, the x-axis is truncated at the 99% quantile value to discount outliers except inventory. For the inventory plot, all points are shown to demonstrate the effect of the earlier segmentation. The truncation of the 99% quantile results in a monotonically decreasing curve without the bump at the upper end.

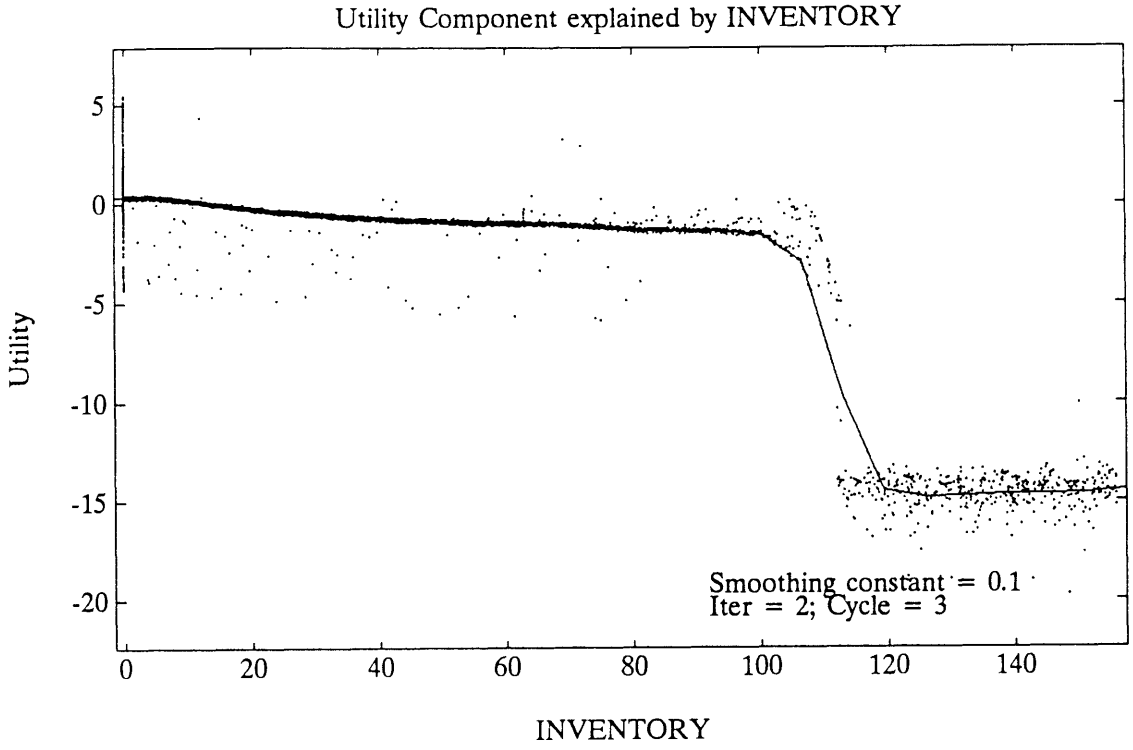


Figure 14: Household inventory utility transformation of the whole choice-based sample by URM when all other variables are linearly specified

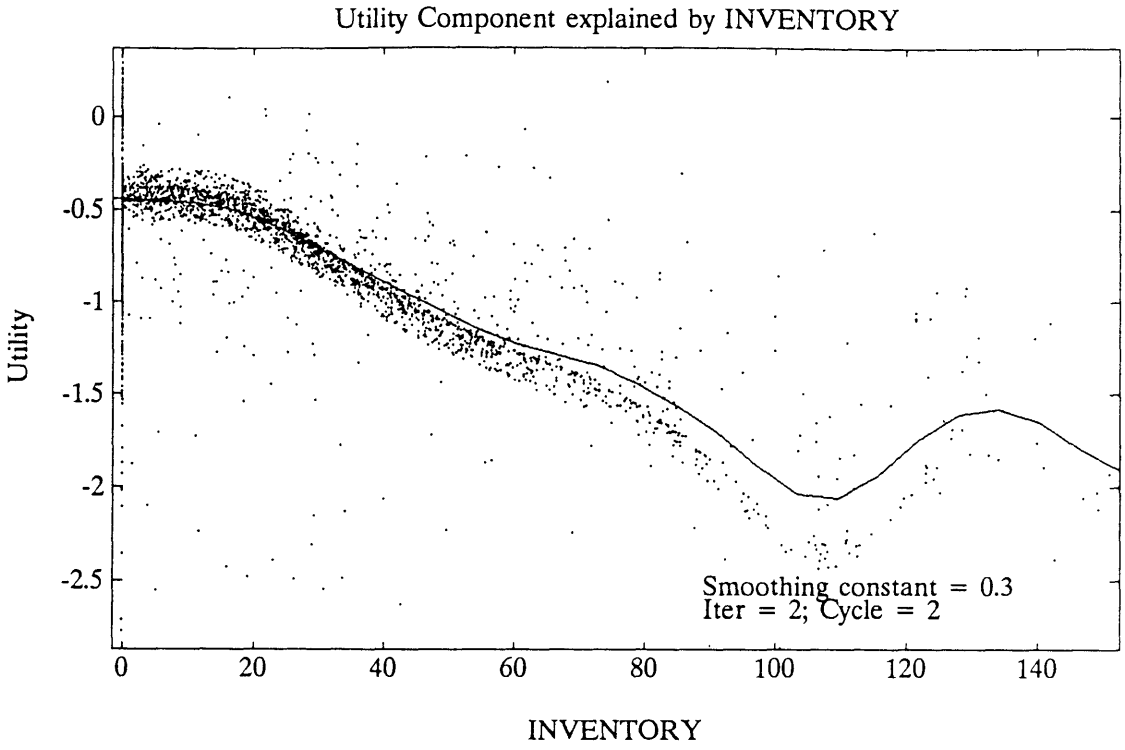


Figure 15: Household inventory utility transformation of the frequent-buyer sample by URM (N=2223)

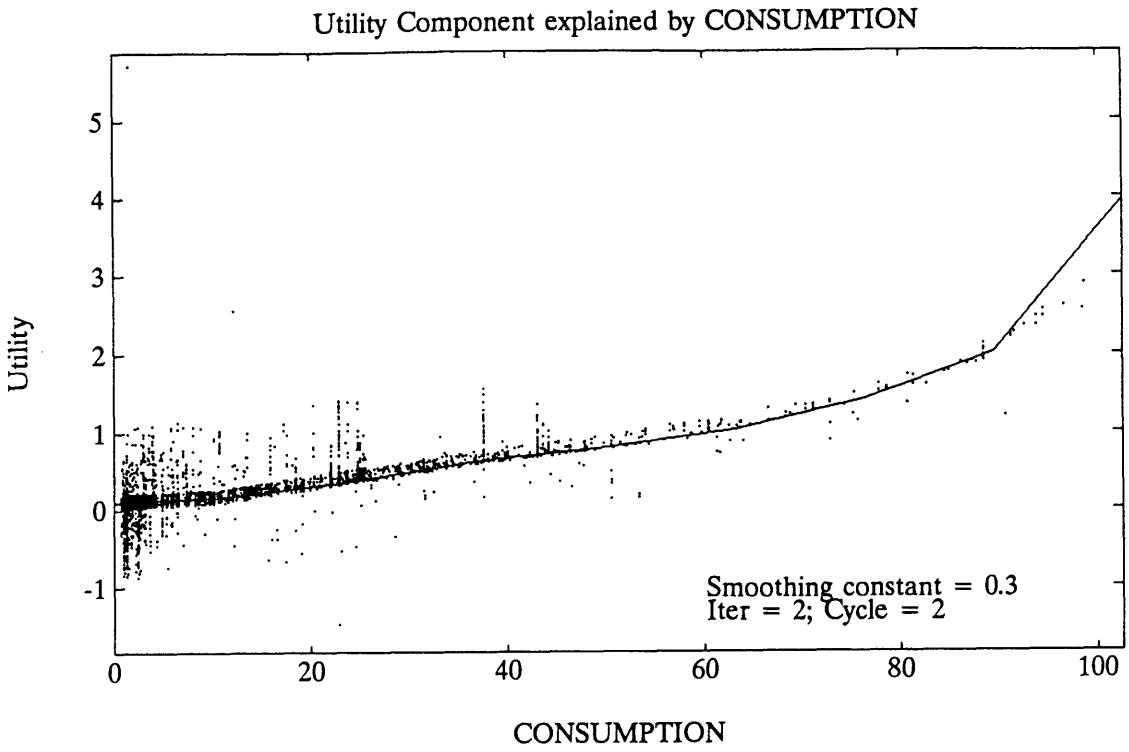


Figure 16: Household consumption utility transformation of the frequent-buyer sample by URM (N=2223)

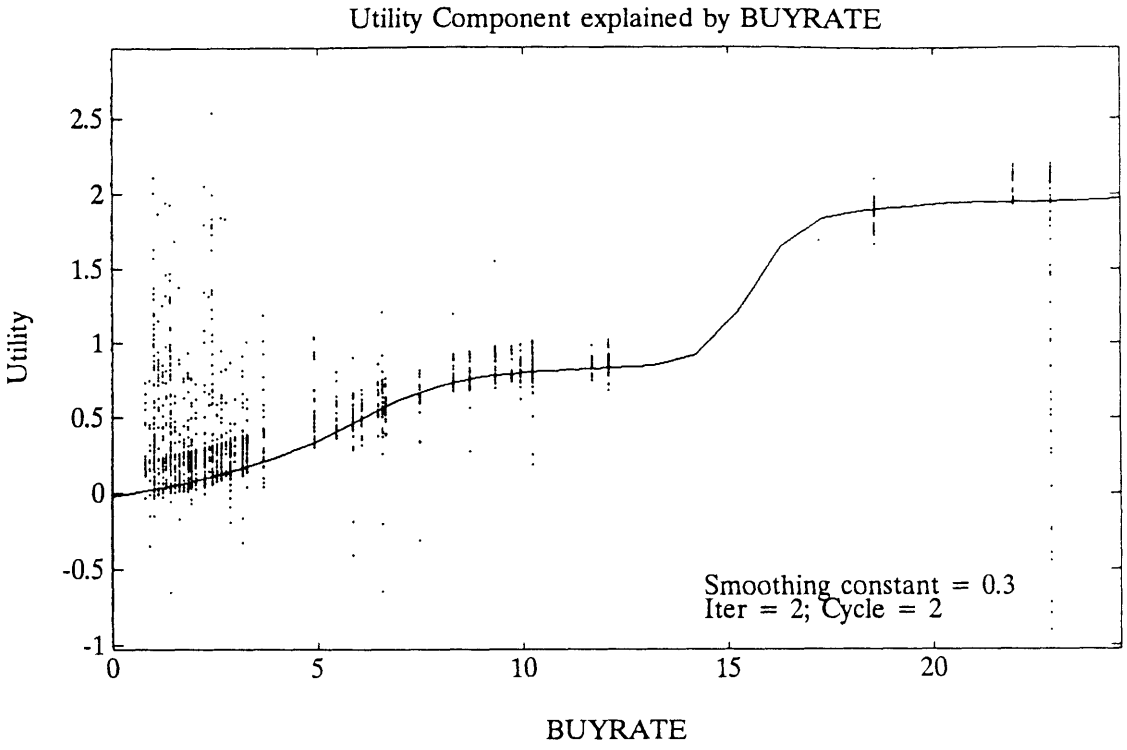


Figure 17: Household buying rate utility transformation of the frequent-buyer sample by URM (N=2223)

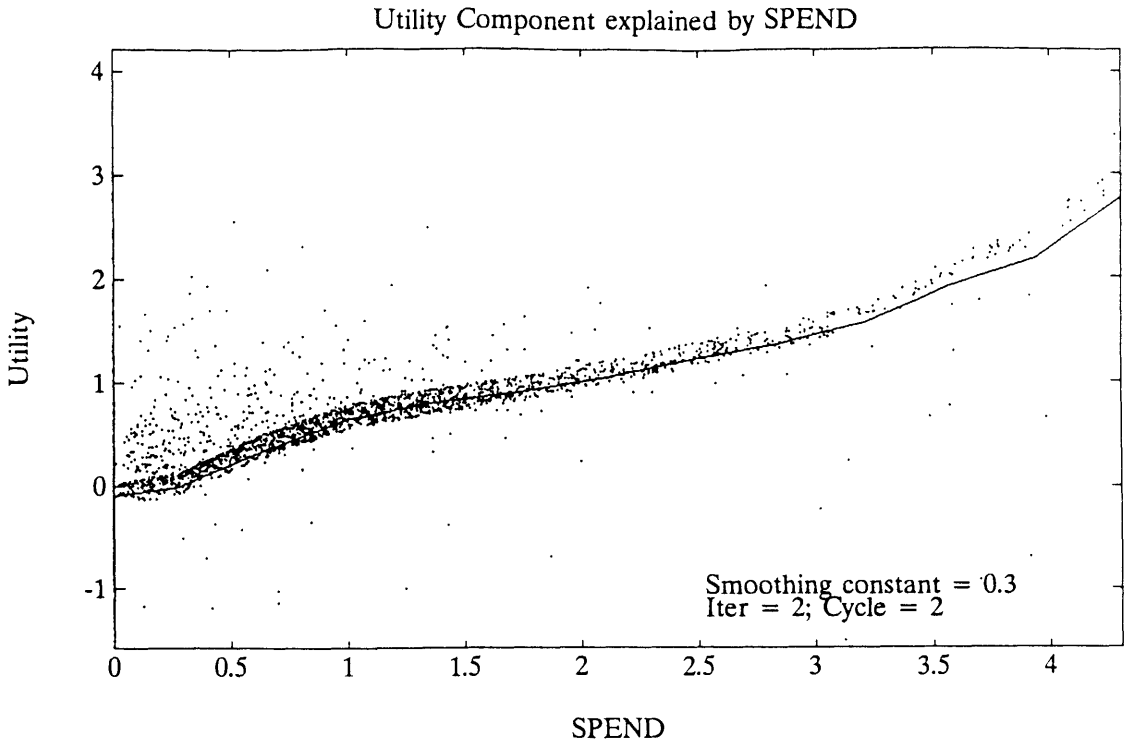


Figure 18: Standardized spending utility transformation of the frequent-buyer sample by URM (N=2223)

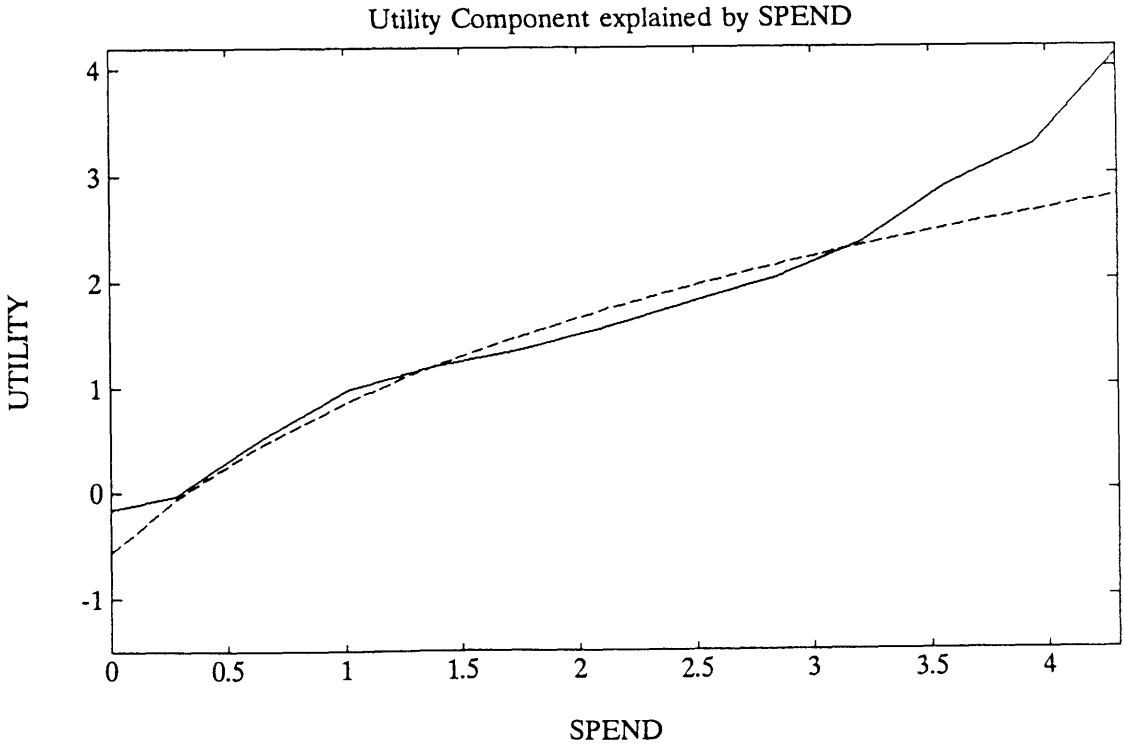


Figure 19: URM vs. parametric transformation of standardized spending variable

5.3 Calibration and Validation

Table 7 illustrates the result of the linear and transformed category purchase model calibrated on the 72-week period. All variables have expected significant signs. Household inventory, standardized spending on a shopping trip, consumption rate, buying rate have the strongest effect followed by category attractiveness, seasonal index, and category attractiveness on multiple unit purchases. The transformed model improves the fit by 0.022 in ρ^2 compared with the linear one.

Table 7: Calibration result of category purchase logit binary model

specification	1	2	3	4	5	6	7	8
$\rho 2$	**0.3852**	**0.3632**	**0.3569**	**0.3540**	**0.3250**	**0.2846**	**0.2102**	**0.1098**
$\bar{\rho}^2$	**0.3764**	**0.3544**	**0.3491**	**0.3473**	**0.3192**	**0.2797**	**0.2064**	**0.1069**
buy-later dummy	3.124 (7.51)	2.256 (5.75)	2.242 (5.62)	1.817 (5.31)	0.929 (2.91)	-0.0227 (-0.078)	-0.543 (-2.02)	0.496 (2.05)
first purchase opportunity	0.963 (2.73)	0.863 (2.44)	0.865 (2.39)	0.445 (1.48)	0.315 (1.05)	0.0877 (0.31)	-0.255 (-0.97)	-0.307 (-1.24)
category attractiveness	0.469 (3.52)	0.504 (3.88)	0.491 (3.82)	0.587 (4.74)	0.586 (4.78)	0.654 (5.55)	0.760 (6.74)	0.743 (7.07)
inventory	-0.0426 (-10.61)	-0.0425 (-10.75)	-0.0425 (-10.78)	-0.0421 (-10.75)	-0.0528 (-13.98)	-0.0526 (-14.10)	-0.0397 (-11.93)	
consumption rate	0.0431 (8.83)	0.0421 (8.71)	0.0423 (8.74)	0.415 (8.61)	0.0486 (10.01)	0.0493 (10.46)		
standardized spending	-----	0.602 (8.59)	0.608 (8.68)	0.600 (8.59)	0.599 (8.49)			
buying rate	0.0716 (7.57)	0.0729 (7.88)	0.0704 (7.69)	0.069 (7.56)				
catego. attract. on multiple unit	0.973 (2.17)	0.964 (2.15)	1.057 (2.33)					
seasonal index	0.145 (3.28)	0.157 (3.58)						
log(spend+1)	2.053 (10.67)							

*** N = 1,487**

In Table 8, the two models are evaluated in the choice-based holdout sample for their probability of correct choices and average loglikelihood as well as R^2 --- correlation between actual and predicted category purchase share by 4-week period.[19] Prediction in the holdout is done using the actual purchase data. In other words, covariates with carry-over effect --- such as household inventory, consumption rate, and category attractiveness, which depend on purchase history --- are derived from the observed rather than forecasted purchases by the model. Hence, the predictive fit tends to be overestimated because the projection into the holdout period uses its own purchase data. We conducted the test in this way to evaluate specifically the model specification of the category purchase by itself to avoid confounding with either the brand choice or the covariates specification. A formal validation in the holdout sample with the complete forecasting must wait until next section where the combined nested logit model is tested.

Table 8: Goodness-of-fit for linear and transformed model in calibration and holdout sample

model		prob. of correct choice	ave. loglikelihood	R^2
Transformed	*calibration*	0.729	-0.426	0.890
	holdout	0.700	-0.518	0.912
Linear	*calibration*	0.718	-0.441	0.886
	holdout	0.692	-0.521	0.932

[19] A calibrated linear model without the seasonality variable is shown below for a reference. There is a larger drop of R^2 in the holdout than calibration without the seasonality index as expected (see footnote 6), despite the fact that the other two criteria do not exhibit such a deterioration.

model		prob. of correct choice	ave. loglikelihood	R^2
Linear	*calibration*	0.718	-0.441	0.886
	holdout	0.692	-0.521	0.932
Linear	*calibration*	0.715	-0.446	0.832
(no seasonality)	*holdout*	0.689	-0.530	0.636

Neither of the models shows much degradation in fit for the holdout compared with the calibration. In fact, R^2 improves. The transformed model surpasses the linear one in all criteria except R^2 in the holdout. A time series plot of the number of category purchases by 4-week period is shown in Figure 20 for each model. Visual inspection indicates a slight advantage for the transformed model as well. At the same time, however, the linear model again demonstrates its robustness in predictive ability. Here, we adopt the superior transformed utility model in our combined nested logit for further considerations.

5.4 The Combined Model

The brand choice and category purchase model are combined to build a sales model at each purchase opportunity in the choice-based sample as

$$\text{sales}(j) = \text{prob}(j \mid \text{category purchase}) \cdot \text{prob}(\text{category purchase}) \cdot \text{size}_j .$$

where size_j is package volume of brandsize j in ounce. Table 9 illustrates performance of the combined model along with a null model which has a constant purchase probability (share of purchases among purchase opportunities) in category decision and the logit model of Section 3 in brand choice. The average sum of square residuals is a sum of squares of the difference between the actual and predicted volume of the 4-week period divided by the number of purchase opportunities, and R^2 is a correlation between the actual and predicted volume for all brandsizes over the periods. The forecasting is done by a Monte Carlo simulation to construct certain carry-over variables --- loyalty, household inventory and consumption rate. The process is repeated 10 times to stabilize the random sample variations.

Table 9: Goodness-of-fit of the combined model in calibration and holdout sample

model	sample	ave. sum of square residuals	R^2
combined model	*calibration*	763.4	0.632
	holdout	802.3	0.577
null model	*calibration*	920.1	0.528
	holdout	901.4	0.565

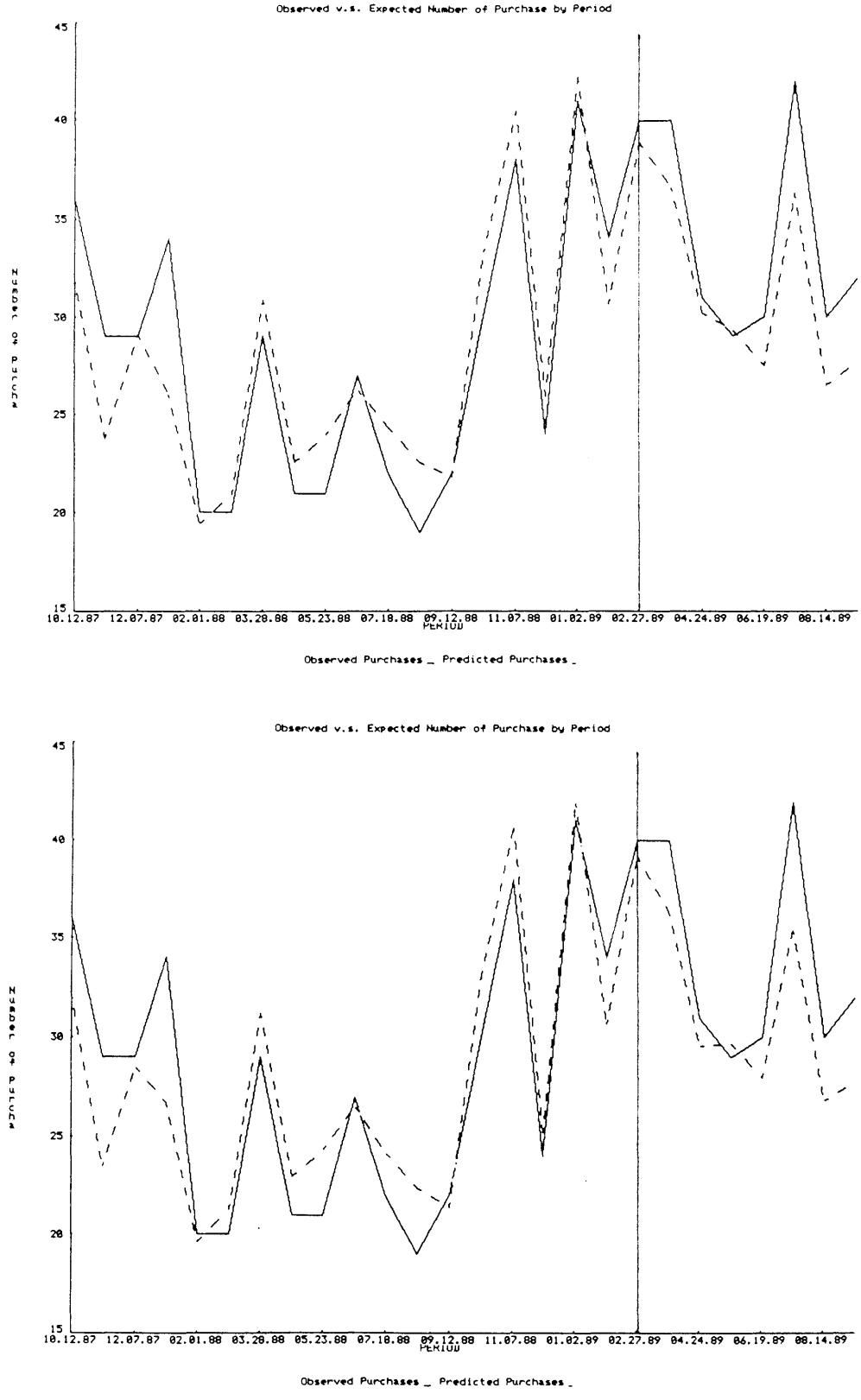


Figure 20: Time series plots of the number of category purchases by linear and transformed model (top: linear, bottom: transformed)

In the holdout, the sum of square residuals exhibits a moderate increase of 5% while R^2 drops by 0.056. The combined model outperforms the null model especially in the sum of square residuals, which is more relevant than R^2 is as far as prediction is concerned. Figures 21 and 22 show the sales tracking of the total Ocean Spray and OS cranberry cocktail 48oz. The result is very encouraging.

6. CONCLUSION

This study has concentrated on a household level marketing mix model which combines the three underlying consumer decision processes --- purchase incidence, brand choice, and quantity selection. A correct model formulation is crucial for understanding and predicting buyer behavior as well as supporting appropriate managerial decision making. In this respect, the nested logit model has advantages for integrating the three interdependent processes, treating purchase quantity, and dealing with shopping trips rather than purchase timing.

The model is calibrated sequentially from brand choice to category purchase, and URM described in Part II is utilized in diagnostics and for inferring parametric utility transformations. The method is shown to be valuable in providing a graphical interface to a model builder for identifying influential points, outliers, and heterogeneous segments, which are otherwise hard to detect in multinomial logit models. The calibrated model is demonstrated to perform well in the cross-validation --- supported by many other marketing applications of logit models in field studies.

The final marketing mix model could provide valuable insights about marketing response relevant to managerial issues. Perhaps, one of the most useful applications is to evaluate effects of marketing mix variables by simulation. In particular, advertising is of great interest. This will be investigated in Part IV.

Figure 21: A time series plot of the total Ocean Spray sales by the combined model

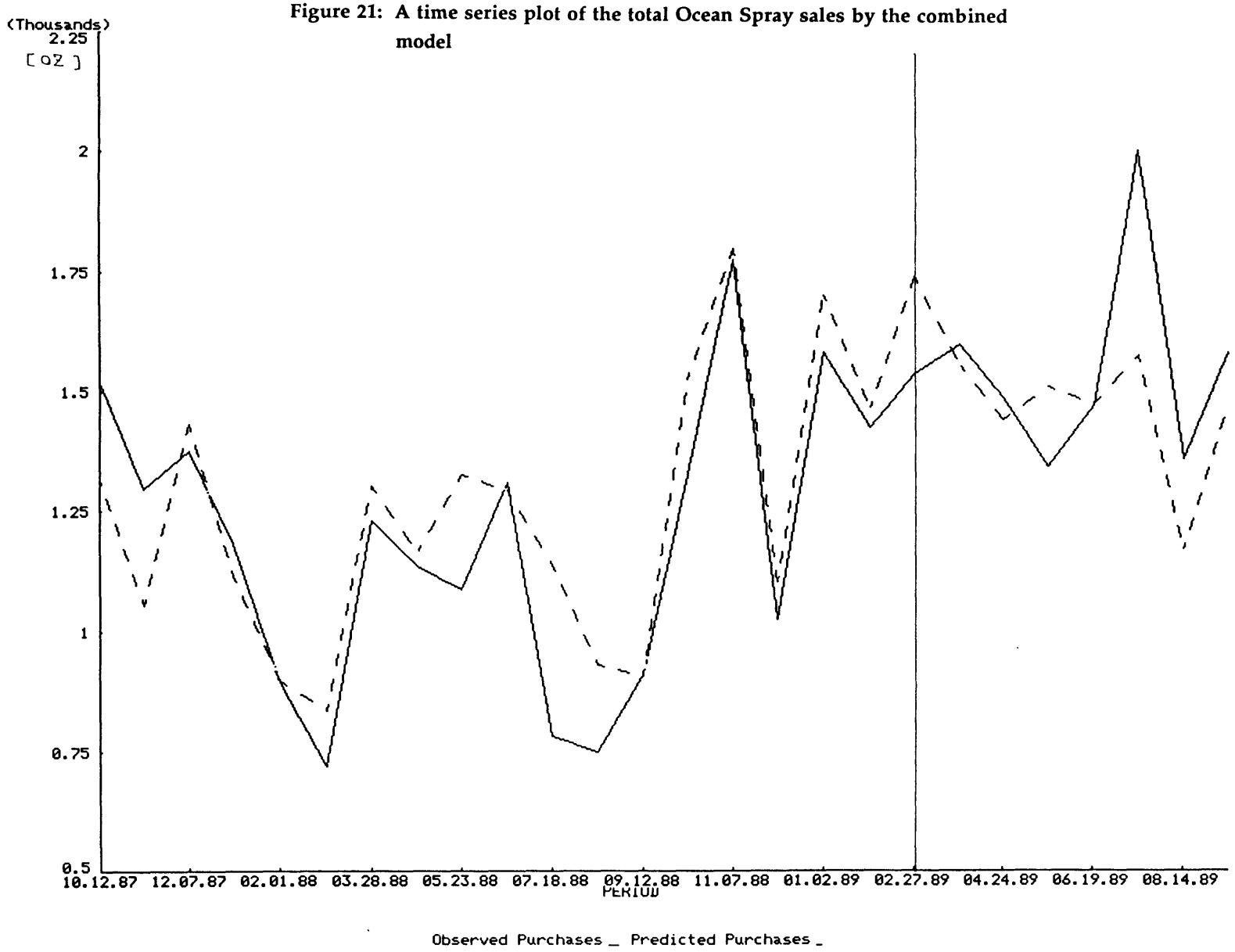
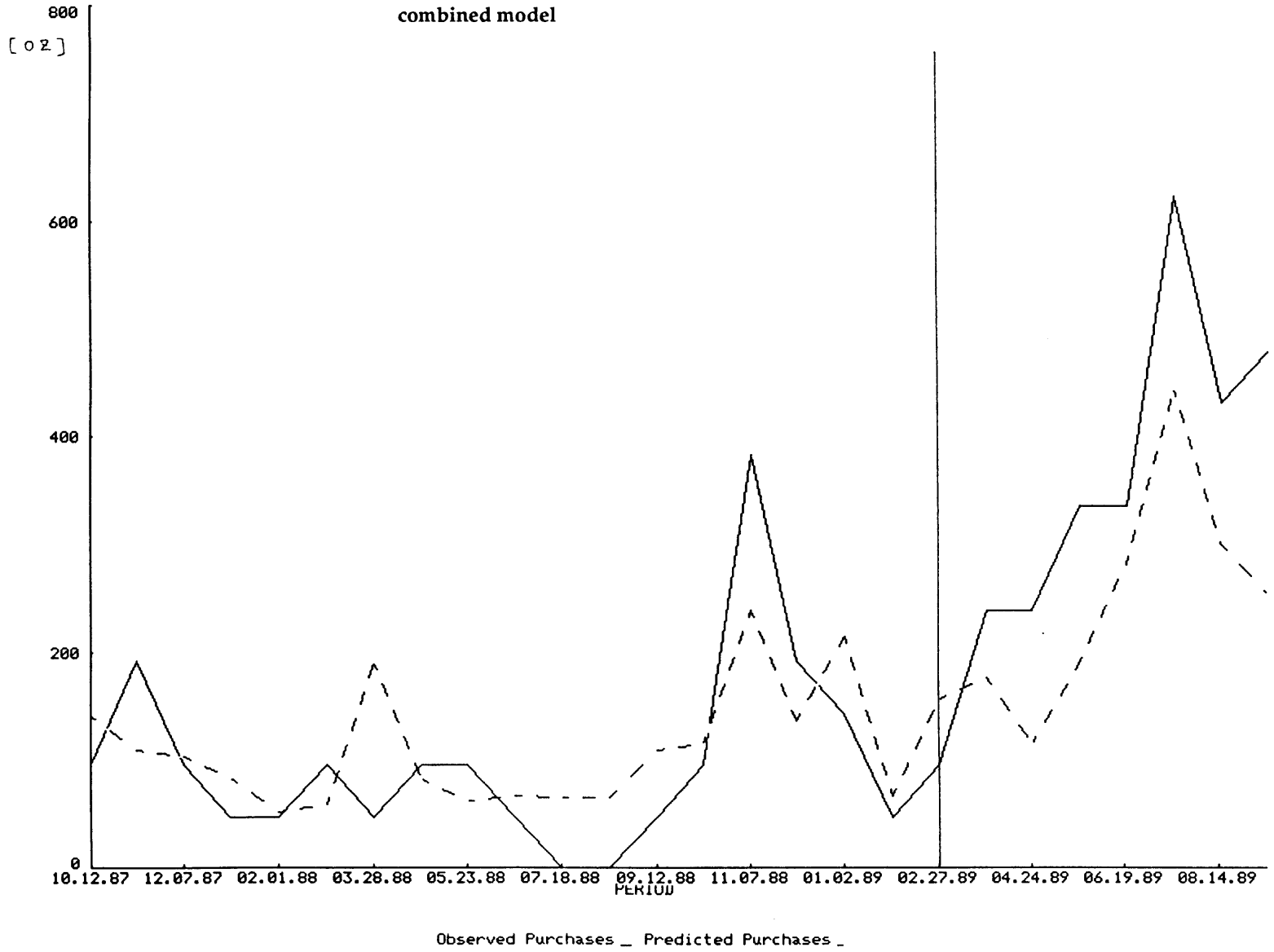


Figure 22: A time series plot of Ocean Spray cranberry 48oz sales by the combined model



TV Advertising Planning Model

Makoto Abe

Operations Research Center
M.I.T.
Cambridge, MA 02139 USA

M. I. T. Doctoral Dissertation, Part IV

June 1991

OVERVIEW

While recent availability of household purchase records and TV ad exposure information from single source data has led to many field studies in buyer behavior, very few have addressed the marketing managers' brand planning problems. This research involves the evaluation of TV advertising plans using a household level marketing mix model that is built from single source data. The model permits a manager to pick alternative advertising scenarios specified by GRPs by week and daypart over a year long period, then computes probability of category purchase and brandsize choice at each shopping trip of the households, and obtains relevant aggregate marketing measures such as brand sales, share as well as average volume consumption and the number of purchases per buyer.

The model consists of two modules. One is a household marketing mix model by the nested logit, which formulates the consumer purchase process by three inter-dependent elements --- category purchase incidence, brand choice, and quantity (size and unit) selection. The explanatory variables are constructed to explain household heterogeneity in brandsize loyalty, inventory, consumption, and purchase timing, as well as carry-over dynamics in the ad effect, consumption, and purchase timing. The second module is a probabilistic exposure model which translates GRP to the household exposure input by accounting for their media habits.

The current study is a result of combining theories from behavioral studies, econometric work for parameter estimation from the data, and stochastic modeling in the ad exposure process. It is found to be capable of including many phenomena heretofore excluded in aggregate advertising models. Various ad scenarios of changing advertising levels and re-allocating GRPs among dayparts are evaluated under illustrative advertising response coefficients to gain insights into the marketing implications of advertising.

1. INTRODUCTION

1.1 Motivation

The availability of large amounts of accurate panel data collected by scanners has prompted many household level studies in marketing modeling. There are many advantages in the disaggregate analyses. One is to avoid much of aggregation bias due to household heterogeneity, and permit deeper insights on buyer behavior.[20] In terms of modeling, one can formulate the whole or parts of a model based on relevant behavioral consideration, which may lead to a more logical and realistic representation and thus better predictions. For instance, in brand sales modeling, a consumer purchase process can be decomposed into three inter-dependent elements --- category purchase, brand choice, and quantity selection, where each one can be modeled by incorporating appropriate buyer decision theories. In contrast, most aggregate sales models are analytical relationships between sales and marketing variables and are typically obtained by regression. As another example, behavioral variables such as purchase-event feedback and advertising forgetting can be readily included in the household level modeling. In aggregate marketing mix models, on the other hand, these phenomena are expressed as a mathematical carry-over function of aggregate sales over time (e.g., lagged terms) for pooled individuals (maybe within a segment) --- which may or may not reflect the underlying individual household behavior.

Thus far, however, much of the disaggregate study has been directed towards modeling the impact of marketing mix variables on buyer behavior and market influence, and very few attempts have been made to address marketing managers' brand planning problems using the scanner panel data (Pedrick & Zufryden 1990). As in the buyer behavioral studies, we could exploit the potential advantages of the household level modeling to better capture dynamic phenomena and act more properly on strategic and tactic issues in managerial decision making.

Furthermore, such studies are not only useful to managers but also to marketing researchers in understanding long-term effect of marketing mix variables, which is rather difficult when just interpreting estimated model parameters. By actually building and running a marketing mix model, it is possible to gain a better insight into the dynamic implications of various marketing activities. A good example is advertising, which is considered to have a relatively weak short-term impact on sales as compared with price and promotion, and whose role may be more

---

[20] With the current scanner data collection method, a panelist consists of multiple members of the household, and this is a limitation in applying behavioral theories. See Kahn, Morrison, & Wright (1986) for the distinction between households and individuals in modelling issues.

long term. As a result, a coefficient for an advertising variable may not exhibit significance even if its carry-over from past exposures are taken into account in the form of goodwill. Sometimes, there exists experimental data (e.g. Behavior Scan[21]) that indicates the effect even when a disaggregate model does not detect a significant advertising coefficient. In such cases, one can justify inclusion of the coefficient on aggregate response grounds.

Some of the reasons for the hesitation towards the disaggregate modeling in the planning may be attributed to complexity in linking managerial control inputs to household inputs. For example, from a manufacture's point of view, household marketing inputs which are monitored by scanners and TV meters --- price paid, store promotions such as feature, display and price-cut, and ad exposures --- must be generated from the manufacture's controllable inputs such as list price, a trade program characterized by allowances and requirements, and advertising GRP or budget. An overall picture of such a complex marketing system is summarized in the diagram below taken from Little (1975) who proposes a fairly complete marketing mix model. For advertising planning, it is necessary to build a model that relates manufacture's media control variables such as GRPs to the household exposures. The current research will focus on TV advertising planning as a first step towards marketing decision making using household level marketing mix models.

1.2 Previous Media Selection Models

The challenge in media selection is to handle a large combinatorial scheduling problem and capture advertising phenomena realistically, while keeping the computation to a reasonable amount. Several different approaches have been proposed. Various mathematical programming methods --- linear, nonlinear, integer, dynamic, and goal programmings --- were introduced in the '60s (Zangwill 1965, Bass & Lonsdale 1966, Little & Lodish 1966, Charnes et. al. 1968, to name a few). Because the objective function and constraints must be expressed in analytic form and then solved, many resorted to some simplifying assumptions such as linear response and no dynamic carry-over in advertising effect. This has led to a lack of realism in such models.

As computational cost continued to decline, simulation models were proposed in the late '60s to '70s (Gensch 1969). They could incorporate a more complex formulation in response functions and exposure distributions since the model needs to be merely evaluated for given inputs rather

[21] Behavior Scan is an experimental market study using scanner data by IRI.

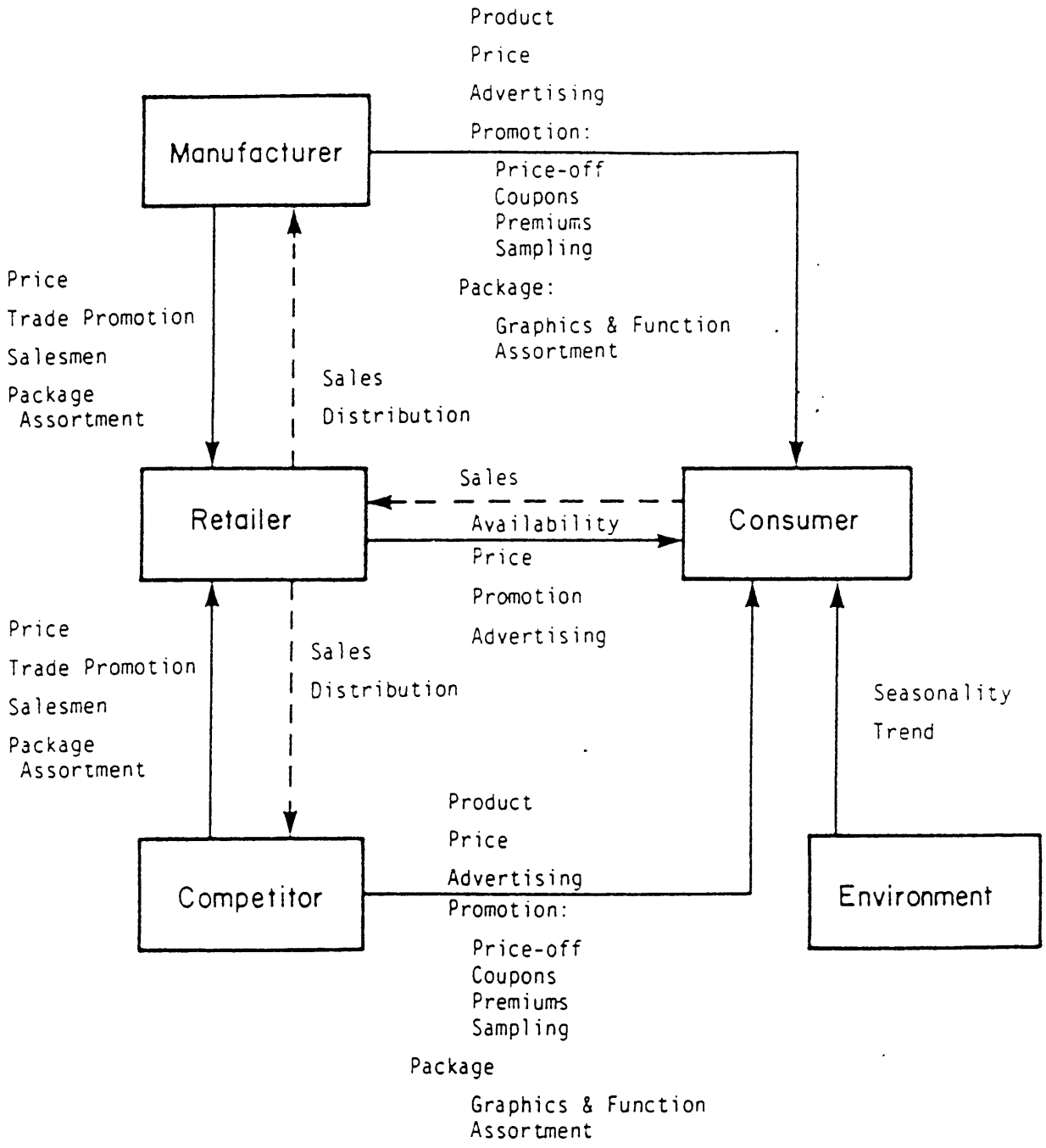


Figure 1: A complete marketing system
(excerpt of Figure 1 from Little 1975)

than to be solved explicitly. The obvious concern of the approach is an enormous amount of computing as the size of the scheduling problem grows.

The third approach is to apply an efficient heuristic combinatorial search method (e.g., greedy and neighbor exchange algorithm) to a data-driven model obtained from statistical and probabilistic analyses (Rust 1985) or to a managerially oriented model by using decision calculus (Little & Lodish 1969, Aaker 1975).

All of the above models operate on aggregate data. To control for consumer heterogeneity, segmentation of target buyers has been commonly practiced. Recently, Pedrick & Zufryden (1990) proposed to evaluate the impact of TV media advertising plans in terms of reach and frequency on a disaggregate marketing mix model calibrated from household level single source data. Although their model offers insight into the reach and frequency trade off, it is not really designed as a dynamic planning model. Nevertheless, it is an important step in a household level media planning approach.

This study investigates media planning of GRP and/or advertising budget by week and daypart to help assess their levels and allocation using scanner panel data. According to one manager, allocation of GRP by daypart is a good starting point for media planning by marketing managers because it is tractable yet does not overly simplify the process. It allows a certain degree of targeting but is not as complex as selecting individual programs, which involves much more effort and is usually handled by ad agencies. The model will permit a manager to generate alternative plans based on his judgement, and then evaluate these scenarios in terms of various marketing measures such as brand sales, share, and profit. In other words, we will construct a household level marketing simulator for advertising managers just like a flight simulator for pilots.

1.3 An Overview of the Model

The model consists of two modules: [1] a household level marketing mix model which accounts for the impact of household marketing mix variables on sales by the three inter-dependent consumer decisions --- category purchase, brand choice, and quantity selection, and [2] a probabilistic model which links manager's advertising control variables to household ad exposures characterized by their media habits.

The first module is a nested logit model introduced and calibrated on the IRI Red Drinks single source data in Part III of this thesis. There, the advantages of the nested logit formulation --- [1] the three decision processes are highly integrated, [2] the method handles discrete and different sizes of the same brand without involving ad-hoc aggregation, and [3] the process is driven by shopping trips --- are discussed in detail.

The second module examines the media habits of each panelist from the observed exposure data and calculates how likely each household is to be exposed to the ads during different time frame of the day referred to as "daypart". Therefore, ad exposures of weekday late night TV watchers will not be much influenced by changes in GRP of weekday mornings, while housewives who watch TV in the morning might be greatly affected. The managerial advertising variable, GRP by week and daypart, is transformed to the household advertising stock variable --- input for the first module --- by incorporating the number and timing of exposures stochastically as well as their memory decay effect.

The features which distinguish the current work from Pedrick & Zufryden (1990) are,

[1] While their managerial advertising variables are reach and frequency over periods in a stationary sense, our control variables are GRP/budget by week and daypart, and the main emphasis is on the dynamic planning reflected in the market movement over a long period of time. In our model, reach and frequency for each week and daypart are treated endogenously. In other words, reach and frequency *within a daypart* cannot be independently manipulated, but are rather a result of interaction between the advertising plan and the model in a systematic way. This is based on the belief that TV programs and household media habits are fairly homogeneous *within a daypart*. Increasing frequency of a daypart is accompanied by increase in reach in some predetermined relationship governed by the underlying stochastic mechanism. Of course, it is still possible to increase reach with relatively small influence in frequency by spreading a given quantity of ads *across dayparts*.

[2] The model operates on the sample households in the database, and marketing measures are constructed by aggregating the sample. Pedrick & Zufryden estimate the population heterogeneity of covariates such as loyalty and ad exposure probability by fitting parametric distributions over the sample, and the aggregate measures are obtained by integrating over the distributions analytically. Obviously, if the generalization of the result to the population is the objective, their approach may seem attractive. The drawback, however, is loss of the advantages in the household level approach and potentially incorrect aggregate estimations by imposing inflexible parametric distributions which are often unimodal even in the presence

of multiple segments. With sufficient number of randomly selected panelists, we can reasonably ensure representativeness in the population at the cost of increased computation.

[3] Their category purchase model does not involve any explanatory variables and is purely stochastic with a Gamma distribution. Hence, expansion or contraction of the total category due to marketing activities is not allowed. The category sales in the nested logit used in our model is affected by marketing mixes --- price, advertising, and promotions --- through the category attractiveness obtained from the brand choice model as well as household characteristics for buying rate, consumption rate, inventory, and expense on a shopping trip.

[4] Our household variables incorporate many carry-over effects, some based on buyer behavior. For example, we explicitly model the forgetting of ad exposures. As another example, both interpurchase time and consumption rate are dynamically constructed to adapt for the household habitual change in the product category. Overall, such an elaborate dynamic model is expected to explain the long-term effect of marketing mix variables better --- an important factor for the market simulator.

The household level marketing mix model, when aggregated, must reproduce phenomena that are considered desirable in aggregate models. Little (1979) identifies five such phenomena in aggregate advertising models.

P1. *Sales respond dynamically upward and downward to increases and decreases of advertising and frequently do so at different rates.*
P2. *Steady-state response can be concave or S-Shaped and will often have positive sales at zero advertising.*
P3. *Competitive advertising affects sales.*
P4. *The dollar effectiveness of advertising can change over time as the result of changes in media, copy, and other factors.*
P5. *Products sometimes respond to increased advertising with a sales increase that falls off even as advertising is held constant.*

The current model accommodates all of them except possibly P5. (We have not actually implemented competitive advertising since there is none in the product category for our sample of households, but the nested logit accommodates the phenomenon.)

The paper is organized as follows. The section 2 first lists key advertising phenomena which need to be addressed, and the two modules are described in reference to them. Section 3 discusses the actual operationalization aspects --- calibration of parameters, an updating

method for the carry-over variables in estimating probabilities of category purchase and brand choice, and performance criteria used for media plans. Then, in section 4, illustrative examples are presented by evaluating examples of advertising plans, and their marketing implications are raised. Finally, section 5 concludes the paper.

2. MODEL

2.1 Objectives

The objective of the current study is to evaluate TV advertising plans in terms of various marketing measures. Here, the advertising plan refers to GRP by week and daypart, or equivalently an advertising budget if the costs of placing ads for each daypart are known. The dayparts divide broadcast hours into six weekday and five weekend segments from early morning to late night, which are shown in Table 1.

Table 1: Time frame of the eleven dayparts

Weekdays (Monday - Friday)		Weekend (Saturday-Sunday)	
1	6 am ~ 9 am		
2	9 am ~ 12 pm	**7**	7 am ~ 1 pm
3	12 pm ~ 4:30 pm	**8**	1 pm ~ 4 pm
4	4:30 pm ~ 7 pm	**9**	4 pm ~ 6 pm
5	7 pm ~ 10 pm	**10**	6 pm ~ 10 pm
6	10 pm ~ 1 am	**11**	10 pm ~ 1 am

As mentioned before, we assume that the dayparts are sufficiently good approximation to capture the household media habits stochastically. That is, the observed exposure data is used to infer average hours per week of TV watch for each daypart by each panelist, that in turn, is used to calculate the expected number of ad exposures received, which are assumed to be uniformly and independently distributed within a daypart.

The marketing measures could be any useful managerial criteria which can be derived from the output of the household marketing mix model --- probability of category purchase and

brandsize choice at each shopping trip[22]. Time series tracking of a particular brand or brandsize sales and share is one of them. By aggregating over a year, common managerial measures such as average brand consumption per buyer, average number of purchases per buyer, and average volume at each purchase occasion, can be obtained. If the cost data for the media and products is available, profit may be computed as well.

What factors and phenomena should be considered in the advertising planning model? We would certainly like to address the four aggregate advertising phenomena, P1 ~ P4 in the last section to our disaggregate household level model. Little & Lodish (1969), in addition, suggest nine essential issues associated with media planning models:

1. *Market segments* for classifying customers
2. *Sales potentials* for each segment
3. *Exposure probabilities* for each media option in each segment
4. *Media cost*
5. *Forgetting* by people exposed to advertising
6. *Seasonality* in product potential and media audience
7. *Individual response* to exposure, including the effect of diminishing returns.
8. *The distribution of exposures* over people and over time
9. *Exposure value* for the exposures in each media option

As we will see, the first module — the nested logit marketing mix model developed in Part III of this thesis — takes into account of 1, 2, 5, 6, and 7. The second module must be designed to represent the probabilistic exposure process of 3, 8, and 9. The only remaining issue, 4, can be readily incorporated into the model if the data is available.

2.2 The Nested Logit Model of Marketing Mixes

Details of the current nested logit marketing mix model are described in Part III of this thesis. Because the nested logit is a household level model, we do not need to classify heterogeneous buyers into separate segments indicated in issues 1, 2, and 7 above. In fact, the advantage of the household modeling is that it can accommodate the household heterogeneity in many aspects

---

22 To be more precise, it is actually purchase opportunity which takes into account of multiple unit purchases. See Part III of this thesis and Guadagni & Little (1987) for detail.

such that, if segmentation were to be done in aggregate models, each cell, consisting of a particular level of each aspect, would contain too few panelists to be analyzed.

The nested logit model employed in this work captures heterogeneity in [1] category purchase via household inventory, buying rate, consumption rate, interpurchase timing, and category attractiveness, [2] brand choice via brandsize loyalty, and [3] ad exposures by a Poisson process with household specific media habit parameters (to be illustrated in the next section).

Forgetting in 5 is modeled in the definition of the adstock, which is a sum of all previous exposures encountered before a particular shopping trip in question, adjusted for memory recall by daily decay. The construct is based on behavioral and field studies of advertising effect (Little and Lodish 1969, Lodish 1971, Clarke 1976, Craig, Sternthal, and Leavitt 1976). The decay rate was estimated to be 0.914 per day from the data in Part III.

In addition to the direct carry-over of advertising by the adstock variable, the model indirectly captures the advertising long-term effect through purchase event feedback variables such as brandsize loyalty in the brand choice and time-varying household consumption rate in the category incidence. All these dynamic considerations should lead to Little's P1 --- the downward sales response due to reduction in advertising intensity is slower than the upward movement from its increase --- because purchase event feedback has a longer time constant than advertising forgetting.

Seasonality in product potential referred to in issue 6 is included in the category model on a weekly basis, which accounts for sales surges during certain holidays like the Christmas and Thanksgiving rather well.

Interestingly, the disaggregate logit model sheds light on the controversy of whether the sales response curve exhibits diminishing return or is S-shaped. When a choice probability of an observation (purchase opportunity or purchase occasion) is plotted against the adstock variable with other variables being fixed, the curve depicts the S-shaped logistic function. But because values of the other variables cause the curve to shift either to the right or left, over the valid domain of non-negative adstock, the response can resemble either the S-shaped or diminishing return as shown in Figure 2.

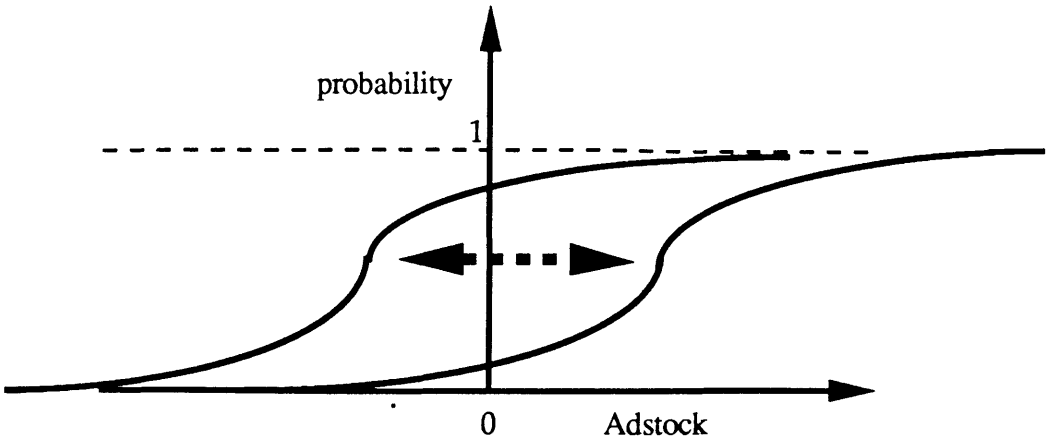


Figure 2: Adstock response curve of choice probability for one observation

Therefore, the advertising response of a particular observation may be either shape depending on the values of the marketing mix variables and household characteristics such as loyalty and inventory at the time of the observation. Furthermore, the choice probability at zero adstock is strictly positive with a varying magnitude affected by other variables, which satisfies Little's P2.

This formulation of the ad response is also appealing from a behavioral point of view. Initial exposures might make a potential buyer aware of the existence of the product, and thus exhibit an increasing return effect.[23] After the initial awareness, the buyer responds positively towards increasing exposures as more cognitive processes such as attention and retention are actively involved. However, repetition of exposures beyond a certain level starts to exploit boredom and saturation in the cognitive process as well as the financial and mental limitation for the product demand. A review of laboratory studies by Sawyer (1981) supports such a diminishing return effect at a high level of the repetition.

The shifting caused by other variables --- especially strong influence from loyalty --- implies that the important moderators of ad response are [1] buyer's prior disposition, and [2] to a lesser extent, other marketing activities such as price and promotions. This is supported by both

[23] Because most laboratory studies measure brand attitudes such as affective and favorable feeling towards the product, they are unlikely to find this phenomenon at a low exposure level, despite the fact that some aggregate analyses support it.

behavioral studies (Cacioppo & Petty 1985, Sawyer 1981) and field studies (Tellis 1988, Kanetker et. al. 1989)

Because the ad response can differ even within an individual due to the situational factors such as marketing activities and household brandsize loyalty and inventory, when the sales response is obtained by aggregating many observations from many individuals, it could produce a mixed result in the response shape that is characterized by the particular market condition at the time --- a combination of product category, and its marketing activities and buyer constituents. Therefore, it does not seem surprising that for the aggregate advertising models some observe the S-shape (Rao & Miller 1975) while others don't (Simon 1969).

Finally, when aggregated over heterogeneous buyers, the change in ad responsiveness due to the shifting of the individual S-curve will manifest itself as a complicated form of interaction between advertising and other marketing mix variables --- which is also addressed in Little (1979)[24].

2.3 Probabilistic Exposure Model

We now turn our attention to the second module --- the probabilistic exposure model --- which transforms GRP by week and daypart to the adstock variable, thereby accounting for the issues 3, 8, and 9.

The adstock of a household at date s, a(s), can be expressed by that of date t_0 as,

(1) $$a(s) = e^{-\lambda(s-t_0)}\, a(t_0) + \sum_k e^{-\lambda(s-t_k)}$$

where k is an index for ad exposures received between date t_0 and s, and t_k is a date of the k-th exposure.

Here, each exposure contributes one unit of effectiveness (could be different for different commercials) to the adstock immediately afterwards, and $e^{-\lambda}$ is a decay rate of the adstock per day, which is estimated to be 0.914 in our data by the method of Fader, Lattin, & Little (1990).

---

[24] Even within the logit model corresponding to one particular observation, interaction among covariates is implicitly being assumed because of the additive utility in the exponential argument.

One way to obtain a(s) is to actually simulate the arrivals of exposures by the Monte Carlo with the underlying probability distribution. However, because the process must be repeated many times to stabilize the outcome for each daypart and panelist, the computational time could be prohibitively large. Hence, we introduce a new analytic methods called the first order Taylor series simulation, where the Monte Carlo simulation of random variables is replaced by their expectations. The logic behind is a result of expanding the random variable around its expectation and keeping up to the first order term.[25] With this method, the second term of (1) --- an increase in adstock due to new exposures --- is substituted by its expectation value.

As mentioned earlier, the advertising exposures are modeled by a Poisson process with a parameter denoted μ'_{hd}, specific to a household-daypart pair, (h,d). Assuming that the daypart sufficiently captures the household media habits, which are considered to be stationary, the exposure rate of panelist h for daypart d of week w, μ_{hdw}, is a linear function of GRP for daypart d of week w, GRP_{dw}, such that

(2) $$\mu_{hdw} = \mu'_{hd} \times GRP_{dw}$$

In other words, doubling GRP_{dw} within a daypart implies twice the exposure rate. μ'_{hd} for each (h,d) pair is estimated from the actual exposure data as

(3) $$\mu'_{hd} = \frac{N_{hd}}{EGRP_d}$$

where $EGRP_d \equiv$ sample mean of weekly GRP_d during all ad flights

and $N_{hd} \equiv$ mean no. of exposures per day for daypart d by panelist h during all flights.

Thus, (2) becomes

(4) $$\mu_{hdw} = \frac{N_{hd}}{EGRP_d} GRP_{dw}.$$

---

[25] Expansion up to the second order involves variance-covariance matrix of the random numbers and is much more complicated and computationally intensive than the first order.

Before describing the construction of the adstock variable, let us derive formulas for two commonly used exposure measures, reach and frequency, for a given weekly GRP level. Reach is defined as a percentage of households receiving at least one exposure, and frequency is an average number of exposures for the households who are exposed to ads. Thus, a product of weekly reach and frequency is equal to weekly GRP. For daypart d of week w, the probability that household h does not receive any exposure is

$$P^0_{hdw} = e^{-\mu_{hdw} D_d} \qquad \text{where } D_d = \begin{cases} 5 & \text{if } d=1\sim6 \\ 2 & \text{if } d=7\sim11 \end{cases}$$

from the Poisson distribution. Because a sum of P^0_{hdw} over households is the expected number of non-exposed, reach for (d,w) is expressed as

$$\text{(5)} \qquad \text{Reach}_{dw} = \frac{H - \sum_h P^0_{hdw}}{H} \times 100$$

where H is the total number of households in the database

Frequency_{dw} is simply $\text{GRP}_{dw}/\text{Reach}_{dw}$.

Similarly, probability of no exposures by household h for week w, P^0_{hw} , is

$$P^0_{hw} = \text{prob (no exposure in weekdays)} \times \text{prob (no exposure in weekend)}$$
$$= \exp\left(- \sum_{d=1}^{n} D_d \mu_{hdw} \right) .$$

Thus, reach for week w is

$$\text{(6)} \qquad \text{Reach}_w = \frac{H - \sum_h P^0_{hw}}{H} \times 100$$

Figures 3 and 4 compare the observed reach and frequency for each week with those of the probabilistic formula of equation (6), and their scatter plots are presented in Figure 5. In all cases, the probabilistic formulation closely approximates the actual reach and frequency.

For the adstock formula, let us suppress the subscripts, h, d, and w for clarity. The expected increase in adstock between t_0 and s due to new exposures whose process is Poisson with exposure rate μ is,

$$E\left[\sum_k e^{-\lambda(s-t_k)}\right] = \int_{t_0}^{s} e^{-\lambda(s-t)}\,\mu\,dt$$

$$= \frac{\mu}{\lambda}\left(1 - e^{-\lambda(s-t_0)}\right) \tag{7}$$

The expected adstock at the end of week i of household h consists of a sum of all expected weekly adstock received prior to week i adjusted for the decay and the expected additional adstock during week i over all dayparts. The operation is graphically summarized in Figure 6, where the distinction between dayparts for weekdays and weekend is made clear.

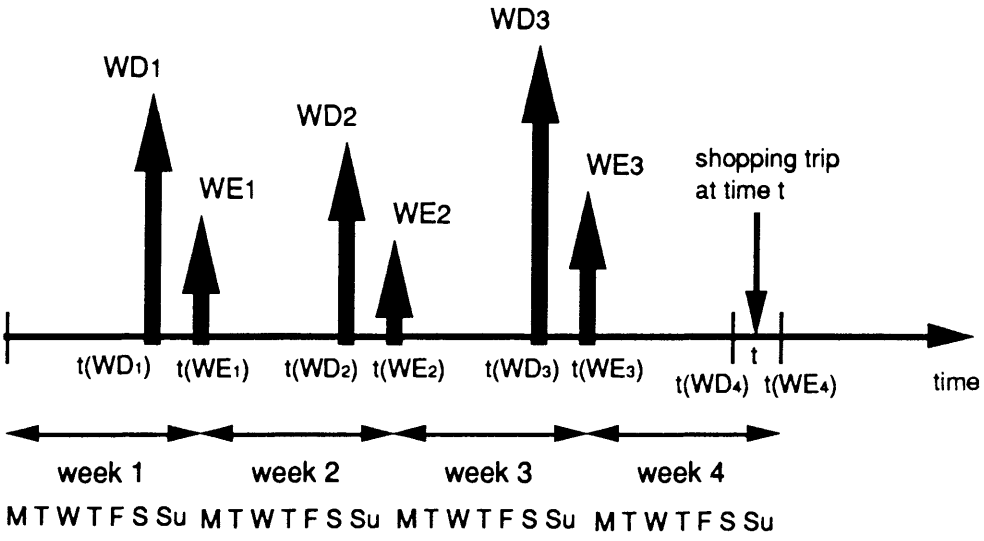
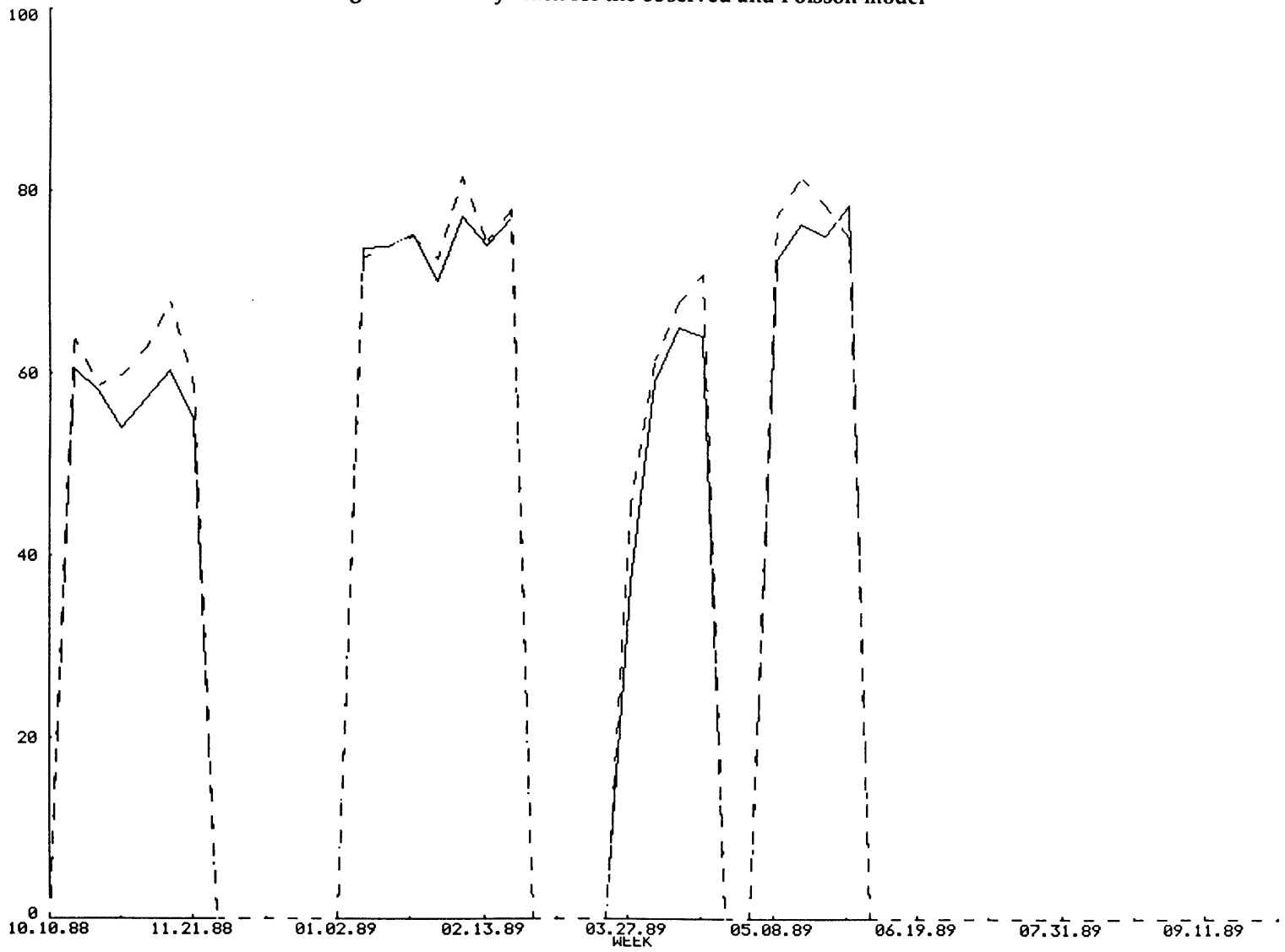


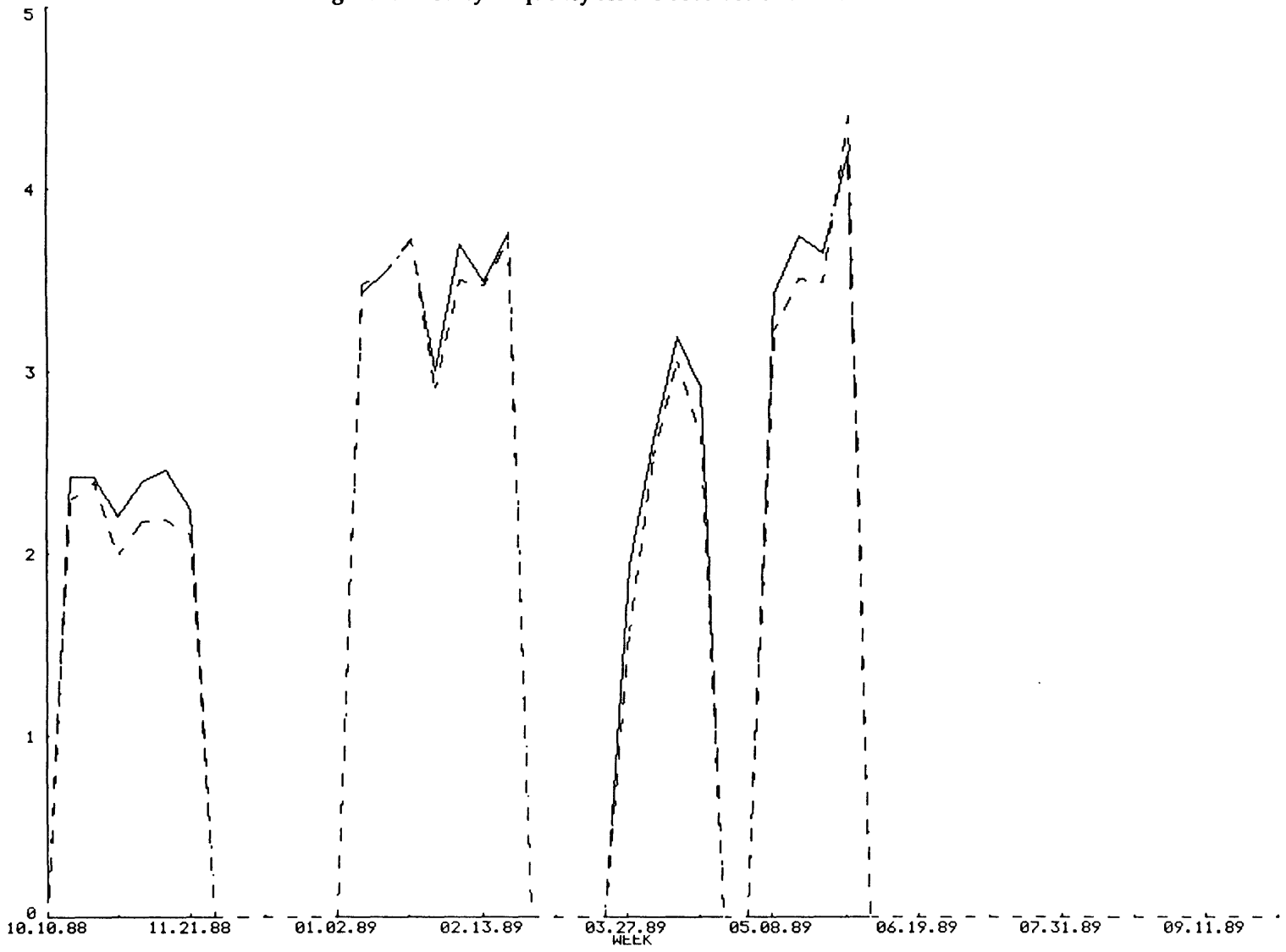
Figure 6 Construction of adstock variable — Graphical Explanation

Figure 3: Weekly reach for the observed and Poisson model



solid = Poisson dash = observed

Figure 4: Weekly frequency for the observed and Poisson model



solid = Poisson dash = observed

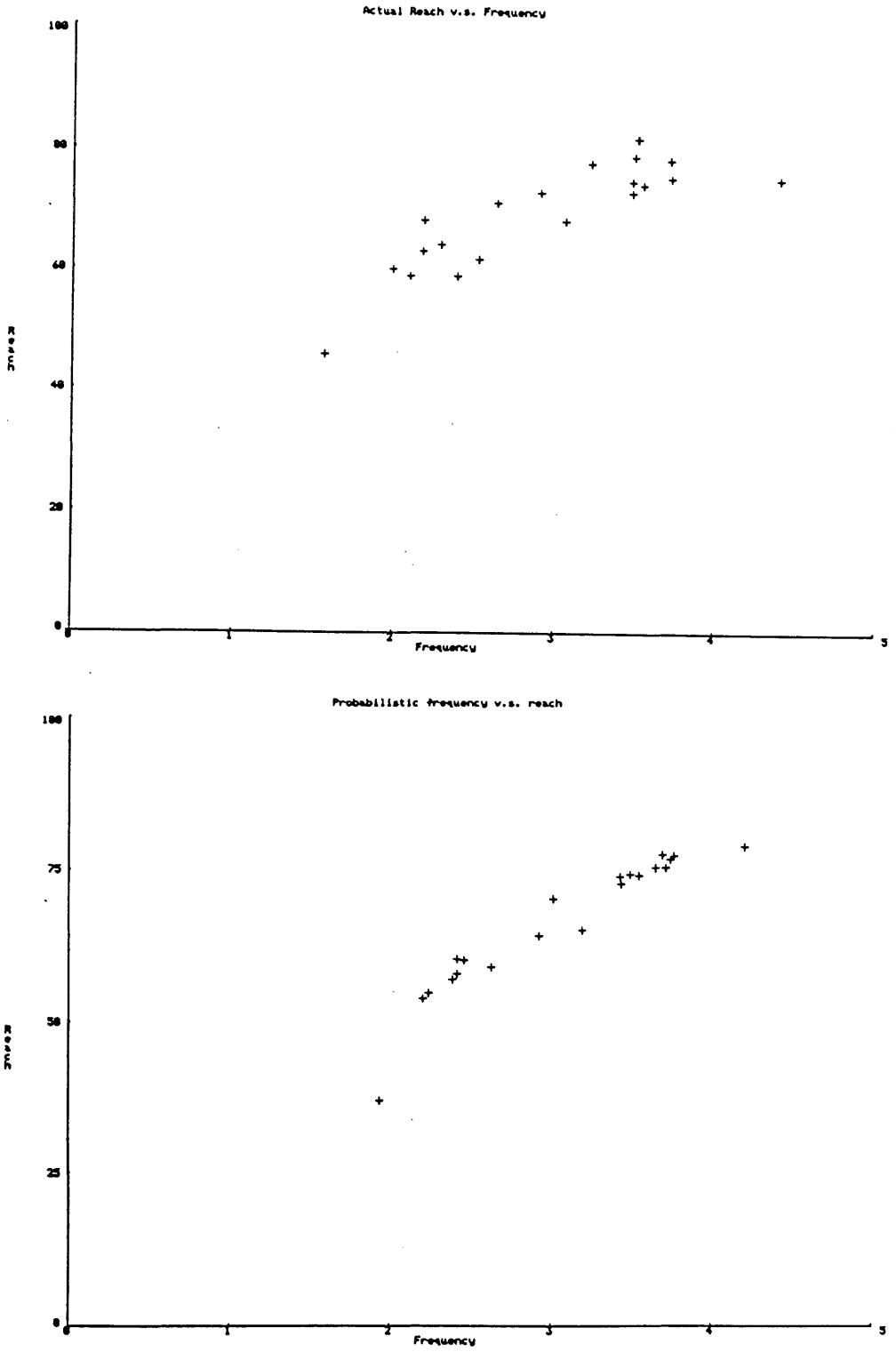


Figure 5: Scatter plots of weekly reach and frequency for the observed and probabilistic formula

The current adstock at date t during week i for daypart d is expressed as

$$a_{hd}(t) = \sum_{k=1}^{i-1} WD_{hdk}\, e^{-\lambda[t - t(WD_k)]} + \sum_{k=1}^{i-1} WE_{hdk}\, e^{-\lambda[t - t(WE_k)]}$$

(8) $$+ \text{E [adstock contribution during week i at date t]}$$

where $t(WD_k)$ = the last date of the weekdays for week k

$t(WE_k)$ = the last date of the weekend for week k

$$WD_{hdk} = E\left[\text{adstock contribution from weekday } GRP_{dk} \text{ at } t = t(WD_k)\right]$$

$$= \frac{\mu_{hdk}}{\lambda}\left(1 - e^{-5\lambda}\right) \qquad d = 1, \ldots, 6$$

$$WE_{hdk} = E\left[\text{adstock contribution from weedend } GRP_{dk} \text{ at } t = t(WE_k)\right]$$

$$= \frac{\mu_{hdk}}{\lambda}\left(1 - e^{-2\lambda}\right) \qquad d = 7, \ldots, 11$$

The last term, E[adstock contribution during week i at date t] --- denoted as NEWAD --- depends on whether date t falls during weekdays or weekend as

Weekdays: $$NEWAD = \frac{\mu_{hdi}}{\lambda}\left\{1 - e^{-\lambda[t-(t(WE_{i-1})+1)]}\right\}$$

Weekend: $$NEWAD = \frac{\mu_{hdi}}{\lambda}\left\{1 - e^{-\lambda[t-(t(WE_{i-1})+6)]}\right\} + WD_{hdi}\, e^{-\lambda[t-(t(WE_{i-1})+5)]}$$

The total adstock is simply a sum of $a_{hd}(t)$ for all dayparts,

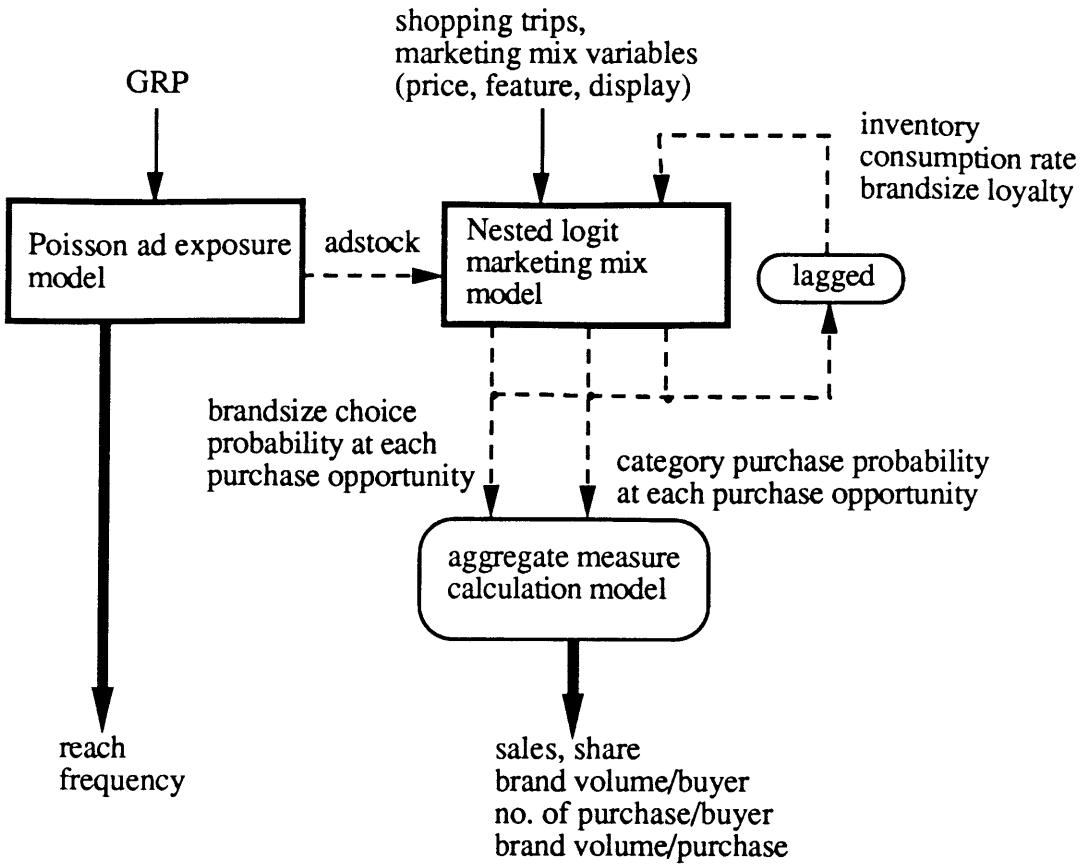
(9) $$a_h(t) = \sum_{d=1}^{n} a_{hd}(t)$$

Equation (9) combined with (8) is the transformation from GRP by week and daypart, GRP_{dw}, to household adstock, $a_h(t)$. The model satisfies the Little's 8 as well as 3 if the terms "media option" and "segment" are substituted by "daypart" and "household" respectively. Because inspection of weekly hours of TV watched during each daypart by households did not reveal

any seasonality in our data, the current formulation does not implement the seasonality of media audience in issue 6. However, the extension is straightforward. By adding another index, p, for the time period (e.g., quarter), a seasonal Poisson parameter, μ'_{hdp}, can be obtained from $EGRP_{dp}$ and N_{hdp} by aggregating those within the period p.

The remaining issue is 9, which refers to different effectiveness among different advertising commercials. Our model can incorporate such effectiveness by varying the unit of the contribution for one exposure in (1) to a value other than 1. If a commercial effectiveness measure is available, say from laboratory studies, it can be readily integrated by scaling up or down the coefficients with an appropriate amount.

To summarize the section, we have discussed the objectives of the current study, and then the model which constitutes of the two modules --- the nested logit household marketing mix model and the probabilistic exposure model --- is proposed. A diagram of the overall model is shown below. The solid arrows corresponds to inputs, the thick arrows to outputs, and the dashed arrows to internal state variables which can be monitored if one wishes. The model takes into account all issues in media planning, 1~9, and aggregate advertising models, P1~P4, suggested by Little.


shopping trips,
marketing mix variables
(price, feature, display)
GRP
inventory
consumption rate
brandsize loyalty
Poisson ad exposure model
adstock
Nested logit marketing mix model
lagged
brandsize choice probability at each purchase opportunity
category purchase probability at each purchase opportunity
aggregate measure calculation model
reach
frequency
sales, share
brand volume/buyer
no. of purchase/buyer
brand volume/purchase

3. OPERATIONALIZATION

3.1 Calibrated Model

The data used in this study is the IRI Red Drinks single source database described in Part III. The variables in the nested logit model are also the same. Exceptions are that all size-specific adstock variables in the brand choice and the adstock in the category purchase are kept in the model, despite the fact that some are not significant. That is because they could have a long-term impact through other carry-over variables even though their short-term effect appears to be weak.

Since the single source database contains the ad exposure information only during the second year, the following procedure is employed. During the first year, the observed purchase data is used to build up the carry-over variables such as brandsize loyalty, household consumption rate and inventory, and then a given advertising plan is simulated in the second year. The coefficients are estimated using the entire two years of the data to achieve the maximum reliability since the model testing was already conducted in Part III.

The variables and their estimated coefficients in the brand choice and category purchase models are listed in Table 2. The buy-later dummy estimated from the choice-based sample is adjusted for the full sample by subtracting log(H/W), where H and W are fractions of the numbers of buy-later observations in the choice-based and entire sample respectively (Manski & McFadden 1981, Chap. 1). The term corrects the undersampling of the buy-later observations in the calibration process by increasing its dummy coefficient. The adstock variable in the category model has been set to be 0.100 for illustrative purposes. This value reflects about 8% sales increase compared with a no advertising case in our database. The similar magnitude was also observed by the aggregate experimental study.

Table 2: Estimated coefficients for the nested logit model

brand choice variables		category purchase variables	
$\rho 2$	0.4597		0.3565
ρ^2-adjusted	0.4519		0.3507
adstock32 **	0.050	buy-later *	5.694 (9.43)
adstock48 **	0.100	first purch. opp.	0.776 (2.60)
adstock64 **	0.050	category attract	0.627 (6.01)
adstock128 **	0.050	inventory	-0.0263 (-9.81)
brandsize loyalty	5.474 (29.85)	consumption rate	0.0296 (8.62)
feature	0.372 (3.29)	buying rate	0.0919 (12.08)
display	1.071 (6.06)	log (spend+1)	1.973 (13.09)
price	-0.985 (-4.77)	catego. attract on multiple units	1.450 (3.56)
		seasonality	0.169 (4.79)
		adstock **	0.100

* The buy-later coefficient is adjusted for choice-based sample by -log(H/W).

** The adstock values have been chosen for an illustrative purpose.

N = 989 for brand choice, and N = 2223 for category purchase model.

Alternative specific constants in the brand choice model are not shown.

Figure 7 shows a histogram of the estimated Poisson exposure parameters for all household and daypart pairs.

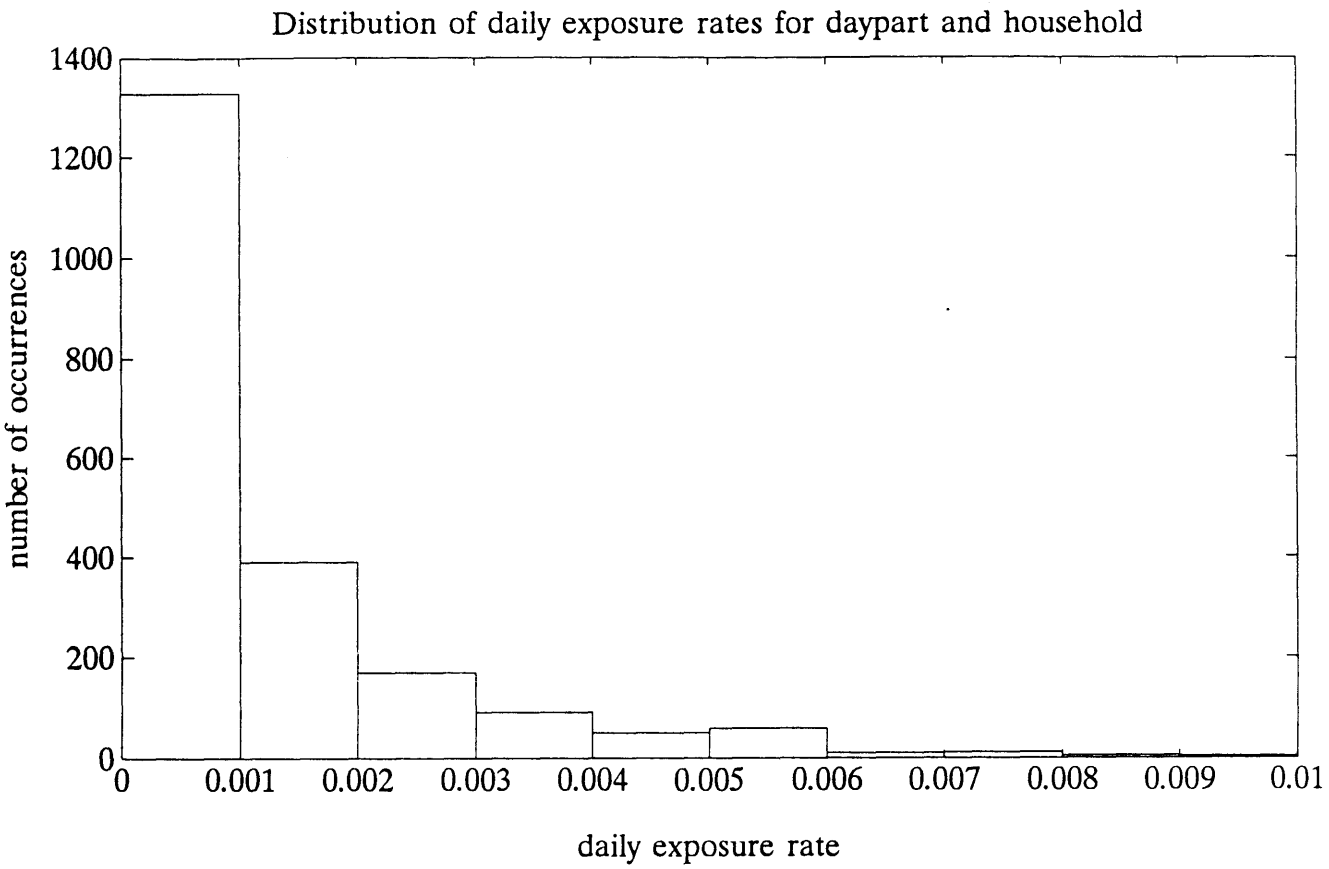


Figure 7: Histogram of the estimated advertising exposure parameters, μ_{dh}

(169 household × 11 dayparts = 1,859 cases)

3.2 Simulation Calculations

Because the model involves certain carry-over variables which require updating based on previous purchases --- brandsize loyalty, household consumption rate, and inventory, --- we must use a purchase sequence predicted by the model as the purchase history to produce tracking of a hypothetical advertising scenario.

Customarily, this has been done by Monte Carlo simulation (Guadagni & Little 1983, 1987, Gupta 1988, Part III of this thesis) with the process being repeated many times to stabilize the outcome. As mentioned in section 2.3, however, a new method is introduced to reduce the computational burden by replacing the simulated random variables by their expectations.

In the following, definitions of the three carry-over variables --- brandsize loyalty, household consumption rate and inventory --- used in Part III are shown.

(10) $\text{loyalty}_j(t+1) = \lambda_1 \cdot \text{loyalty}_j(t) + (1-\lambda_1) \cdot d_j(t)$

where $d_j(t)$ is 1 if alternative j is bought at t-th purchase occasion, 0 otherwise, and λ_1 is estimated to be 0.774 from the data.

(11) smoothed consumption rate (at current category purchase) =

$\lambda_c \cdot$ smoothed consumption rate (at the previous category purchase)

$+ (1-\lambda_c) \cdot$ (volume of the previous purchase / smoothed interpurchase time)

where λ_c is estimated to be 0.49 from the data.

(12) $$\text{inventory(r)} = \text{inventory(r-1)} - \frac{[\text{date(r)} - \text{date(r-1)}]}{7} + \frac{\text{volume purchased (r-1)}}{\text{hh buying rate}}$$

For the brandsize loyalty variable defined in (10), the observed choice variable, $d_j(t)$, is substituted by the predicted probability, $p_j(t)$, to be

(13) $\text{loyalty}_j(t+1) = \lambda_1 \cdot \text{loyalty}_j(t) + (1-\lambda_1) \cdot p_j(t)$

The smoothed interpurchase time used in the consumption rate is adaptively updated based on past purchase history as (14).

(14) smoothed interpurchase time (at current trip) =

$\lambda_t \cdot$ smoothed interpurchase time(at previous category purchase)

$+ (1 - \lambda_t) \cdot$ (most recent interpurchase time) .

where λ_t is estimated to be 0.73 from the data.

Here, the most recent interpurchase time for computing the household interpurchase time (14) is replaced by

(15) $$\text{most recent interpurchase time} = \frac{\text{most recent intertrip time}}{\text{category purchase probability at the trip}}$$

For the household consumption rate shown in (11), volume purchased at the previous purchase is replaced by its expectation at the previous trip if purchase was made as

(16) $$\text{volume purchased (trip)} = \Sigma_j \; p_j(\text{trip}) \times \text{size}_j \times \text{correction}$$

where size_j is package volume of brandsize j in unit of ounce. The correction factor accounts for the possibility of multiple unit purchase in a trip. Since the expected number of unit purchased can be approximated as

E [no. of units purchased]

$$= P_1 (1-P_2) \cdot 1 + P_1 P_2 (1-P_3) \cdot 2 + P_1 P_2 P_3 (1-P_4) \cdot 3 + \cdots$$

$$\cong P_1 (1-P_2) \cdot 1 + P_1 P_2 (1-P_2) \cdot 2 + P_1 P_2 P_2 (1-P_2) \cdot 3 + \cdots$$

$$= P_1 (1-P_2) \cdot (1 + 2 P_2 + 3 P_2^2 + \cdots + k P_2^{k-1} + \cdots)$$

$$= P_1 (1-P_2) \cdot 1 / (1-P_2)^2$$

$$= P_1 / (1-P_2)$$

where P_k denotes the category purchase probability at the k-th purchase occasion of a trip, (P_k's differ because of different values of explanatory variables in first purchase opportunity dummy, attractiveness on multiple unit, and hh inventory even within a trip.[26])

the correction factor in (16) is

$$\text{correction} = 1 / (1-P_2)$$

Finally, volume purchased at the (r-1)st purchase opportunity for the household inventory in (12) is substituted by its probabilistic expression as

(17) $$\text{volume purchased}(r-1) = \Sigma_j \; p_j(r-1) \times \text{size}_j \times p_{cat}(r-1)$$

The decay constant, λ, for loyalty, consumption rate, and interpurchase time shown in (10), (11), and (14) are estimated by the Taylor series method (Fader, Lattin & Little 1990) for each purchase made. In applying these to updating on a shopping trip basis, the values must be adjusted to account for the higher frequency of trips (47,272) than purchases (989). This is done by raising λ to the power of 989/47272 as $\lambda_{trip} = \lambda_{purchase}^{(989/47272)}$.

3.3 Aggregate Marketing Measures

In addition to time series tracking of category sales and brand sales and share, the following aggregate measures over the period of interest (in our case, the entire second year) are computed.

Brand volume

---

[26] The approximation of P_k's for k>2 by P_2 causes a slight overestimation because $P_k < P_2$ for k>2 due to accumulated product inventory from the previous purchases. However, the error is the second order since the approximation is used for estimating the previous volume purchase to update only consumption rate but not inventory which is explicitly updated after each purchase opportunity even within a trip as in (16).

Brand share

Brand volume per buyer

Number of purchase occasions per buyer

Brand volume per purchase occasion

Brand volume per buyer measures an increase in the brand consumption induced by the ads. This can be decomposed into shorter household interpurchase time and purchase of a larger size, which are reflected respectively in the number of purchase occasions per buyer and Brand volume per purchase occasion.

The number of purchase occasions per buyer is expressed as

$$\text{(18)} \qquad \frac{\sum_{h} \sum_{n \in h} p_{cat,h}(n)}{\text{No. of buyers in the sample}}$$

The brand volume per purchase occasion is simply brand volume per buyer divided by purchase occasion per buyer.

4. ILLUSTRATIVE EXAMPLES AND MARKETING IMPLICATIONS

In this section, various ad scenarios are evaluated on the model to gain insights into the advertising effects. All runs are performed on households who have made more than two purchases during the two years of the sample period as done in Part III for the calibration We have randomly selected 20 households to save the computational time. (The same 20 households are used for all runs.) The subsample constitutes 2,398 trips in the testing year.

4.1 Comparison of No Ad, Base Ad Case, and Twice the Base Case

During the second year of the sample period between 10-10-88 and 10-08-89, Ocean Spray ran TV advertising in four separate flights whose lengths ranged from 4 to 7 weeks (Figures 3 and

4). Although our advertising response coefficients are hypothetical, we shall build our advertising scenarios around the actual TV exposures as a base case. The first scenario of interest is a no advertising case. This will help us assess how much the implemented advertising plan would produce in terms of sales and share. The second case is a scenario where GRP for each daypart and week is doubled. The question addressed here is how much could be gained by allocating twice as many GRP?

Figures 8 and 9 show weekly reach and frequency for the base and two times the base GRP respectively using the formula of (6). Also shown are their scatter plots in Figure 10, which clearly indicate the nonlinear relationship between reach and frequency. For instance, doubling the GRP increases frequency by 32% and 72% for the extreme lower left and upper right points respectively, while their reaches increase by 53% and 20%. In other words, boosting GRP within a daypart exhibits the diminishing return effect on reach, which intuitively makes sense because its upper bound is 100. Table 3 is an aggregate summary of the no ad, base ad, and twice the base ad cases over the year. It suggests that, for the advertising response coefficients used here, a large part of the sales increase due to the advertising comes from more frequent purchases, and the contribution from size trade-ups is only one quarter of it.

Table 3: Aggregate summary of no ad, base, and two time the base cases over the 52 weeks

measure	Base	No ad	Twice the base
Category sales [oz]	7,087.7	6,596.9	7,650.2
(% change from base)		(-6.9%)	(+7.9%)
OS sales [oz]	5,177.3	4,740.9	5,704.9
(% change from base)		(-8.4%)	(+10.2%)
OS share [%]	73.05	71.86	74.57
OS vol / buyer [oz]	258.9	237.0	285.25
No. of purchase / buyer	7.31	6.81	7.89
(% change from base)		(-6.9%)	(+7.9%)
OS vol / purchase [oz]	35.4	34.8	36.2
(% change from base)		(-1.7%)	(+2.1%)

Figure 11 is a time series of Ocean Spray sales constructed by aggregating category purchase and brand choice probabilities of trips made by the 20 households for each four weeks. The sales increase during the ad flights is followed by some drop afterwards. The phenomenon can be

Figure 8: Weekly reach for base and twice the base cases

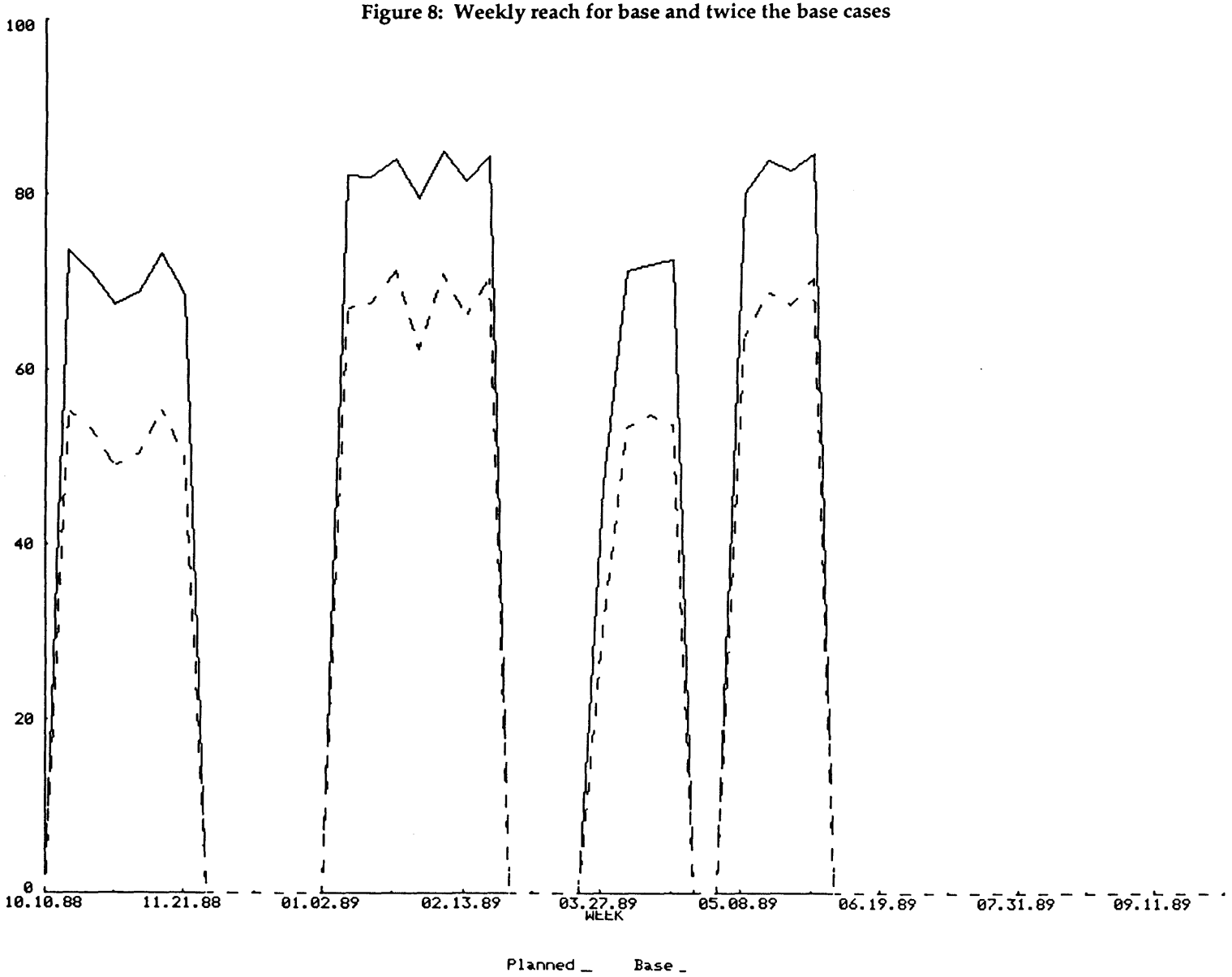
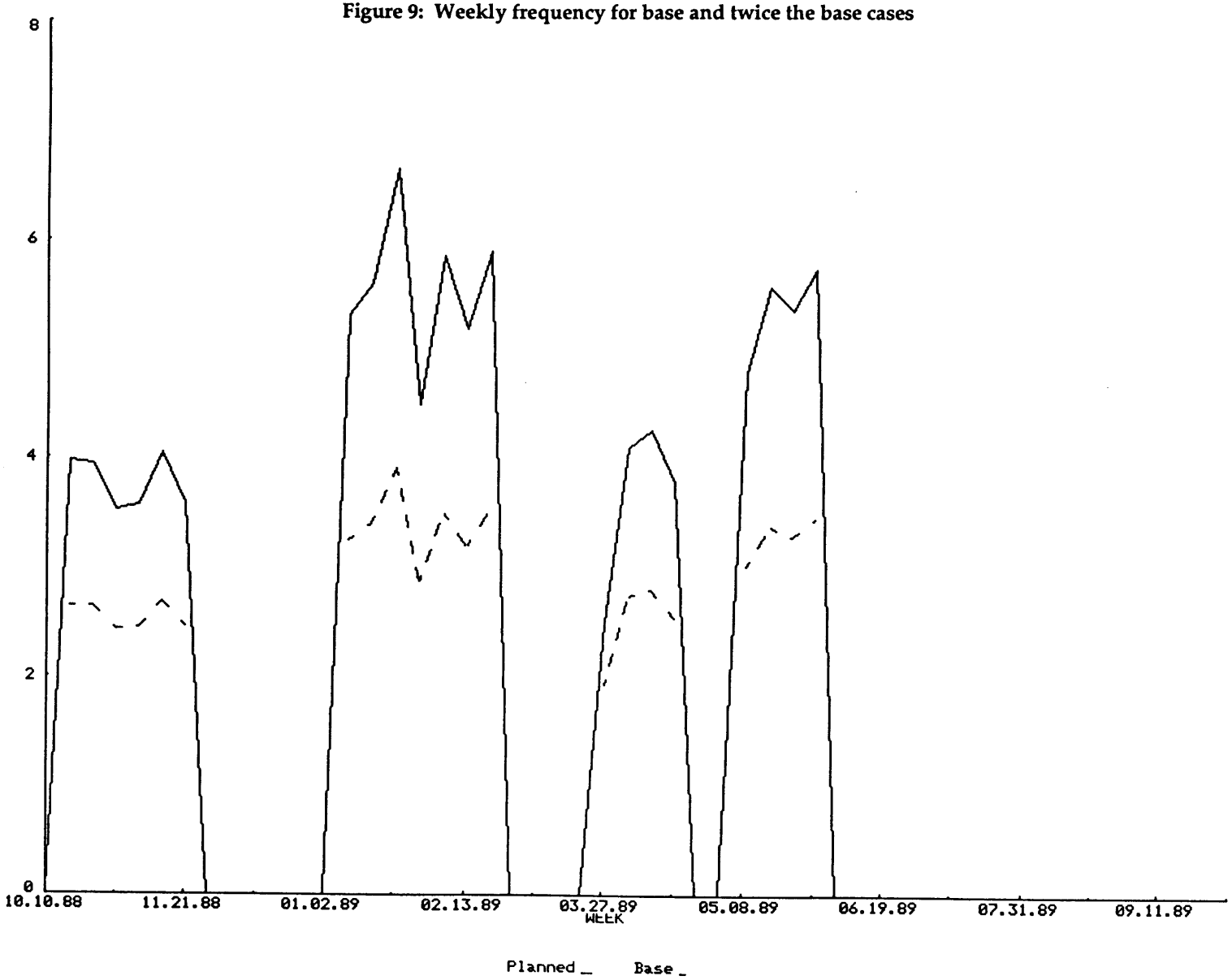


Figure 9: Weekly frequency for base and twice the base cases



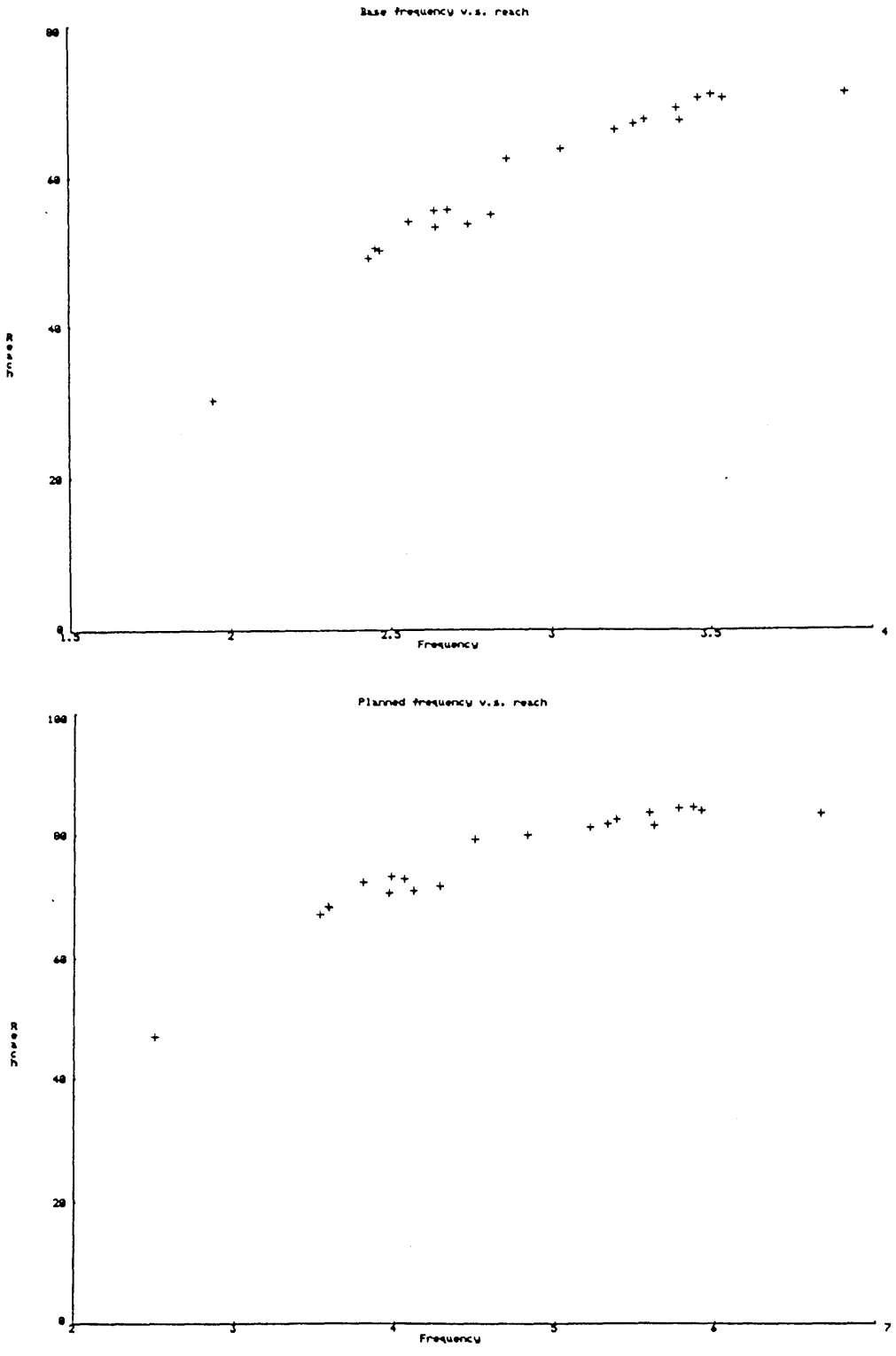
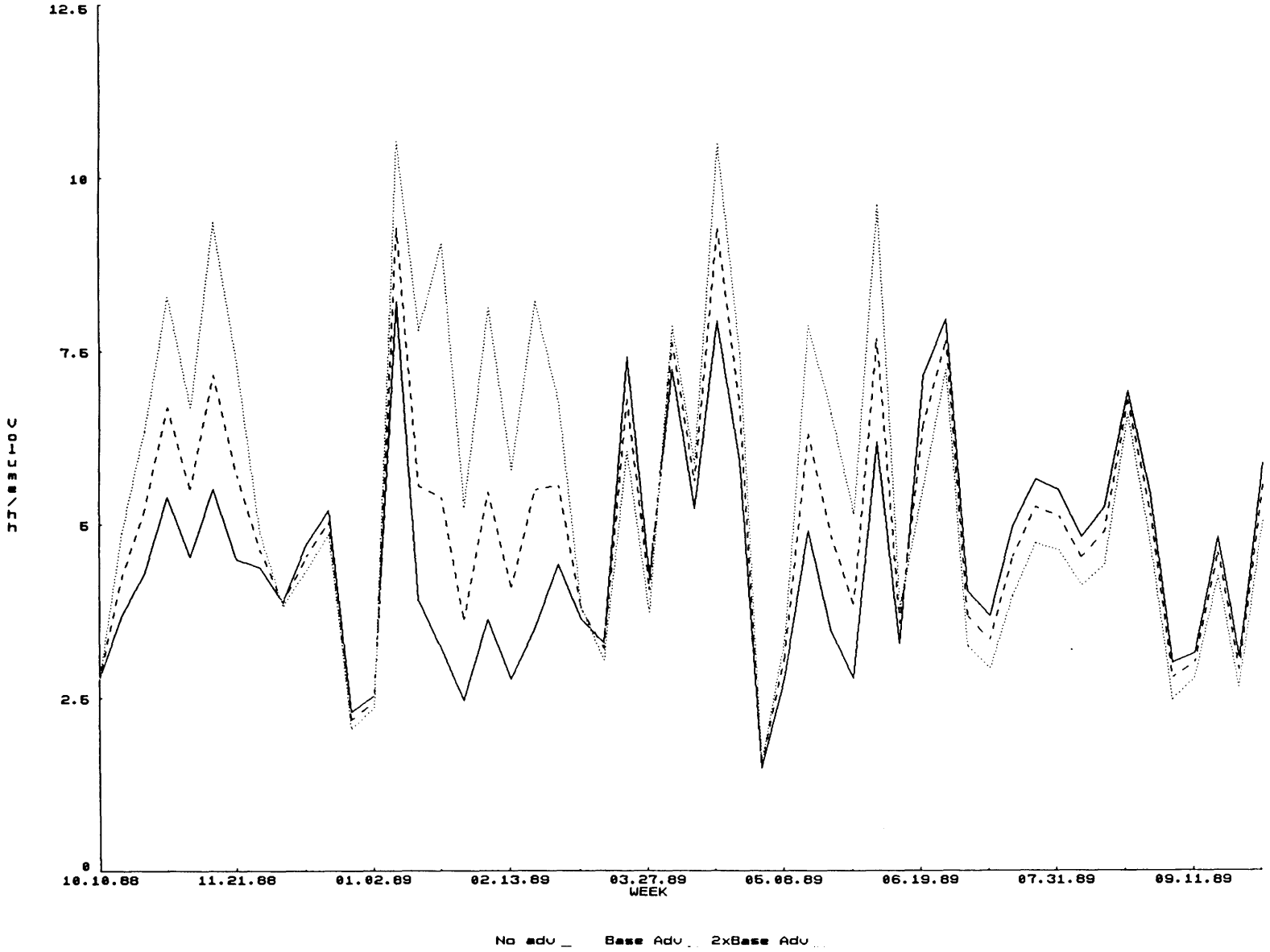


Figure 10: Scatter plots of weekly reach and frequency for base and twice the base cases

Each point corresponds to one week. Note change in scale of the frequency axis between plots

Figure 11: Total Ocean Spray sales for no ad, base, and twice the base cases



described as an advertising induced purchase acceleration, which in turn results in the subsequent low sales because of a high level of inventory. The household inventory is the strongest carry-over variable (t=-9.8) in Table 2. Furthermore, because the consumption rate --- another strong carry-over variable (t=8.6) --- has a large decay constant (λ=0.49), any increase in the consumption by the advertising is temporary and does not have an enough lasting effect to override the sales drop caused by the inventory. In contrast, the share plot shown in Figure 12 does not exhibit such a purchase acceleration phenomenon since the inventory variable appears only in the category purchase model.

Figure 13 plots the annual category and Ocean Spray sales by varying the GRP from 0 to 10 times the base case. The aggregate advertising response functions exhibit nonlinear relationships, however, not exactly the S-shape. For GRP less than 5 times the base level, the plots show a mild increasing return, while a diminishing return is observed at extremely high levels.

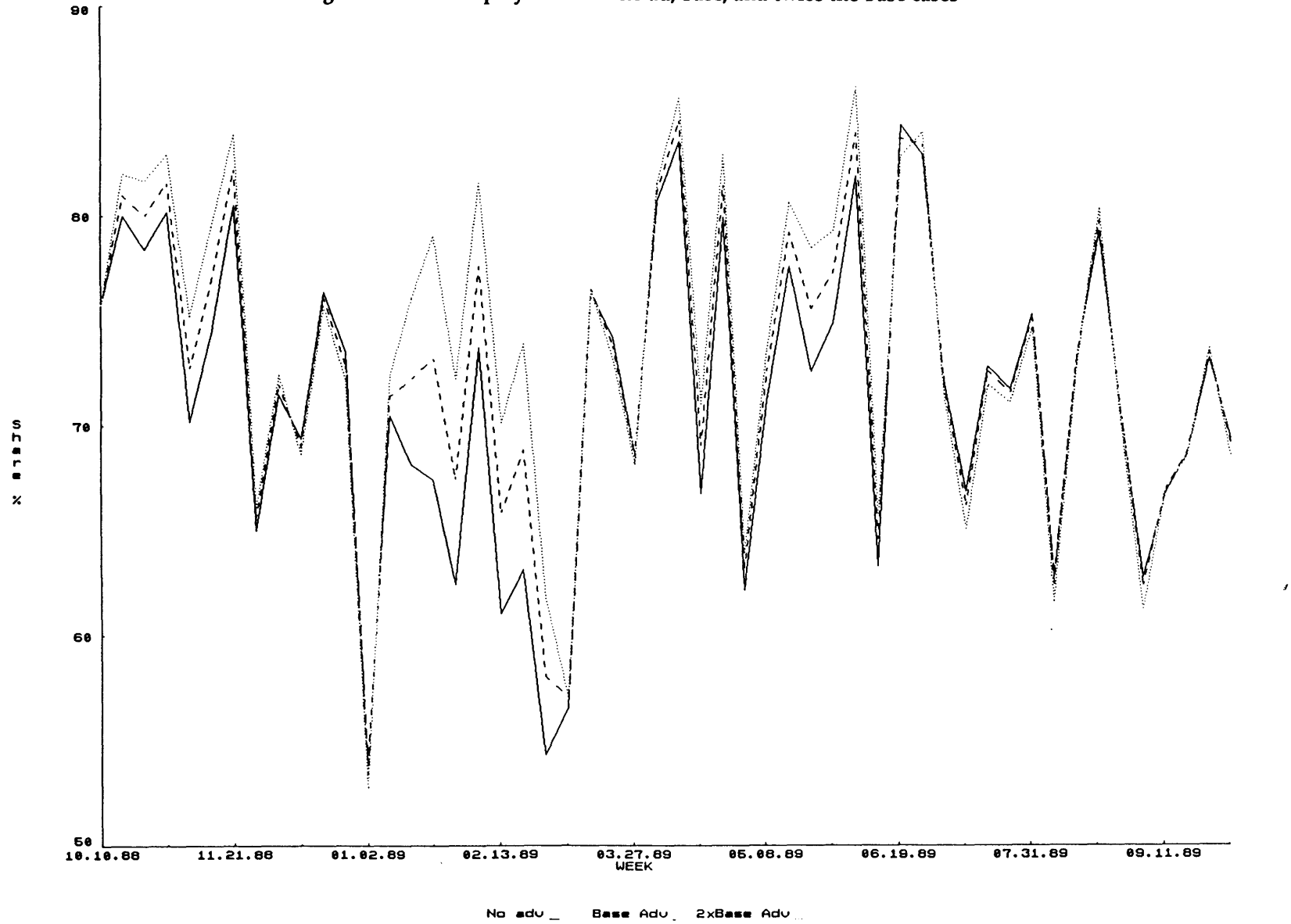
4.2 Re-allocation of GRP among dayparts

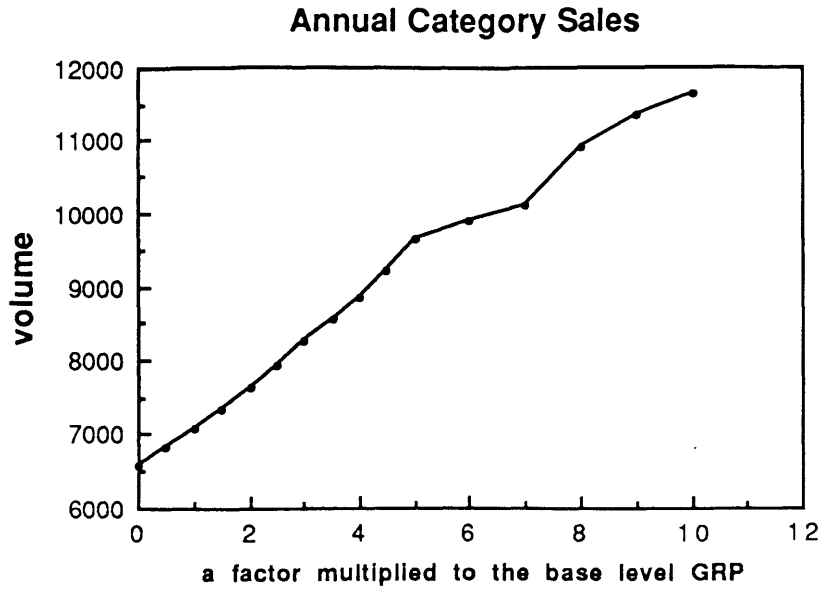
To examine the effectiveness of advertising among different dayparts, the observed weekly GRP is allocated to a single daypart. By concentrating GRP into a single daypart, reach is limited by targeting ads to a segment of households with a particular media habit and a large increase in frequency can be achieved by repeated exposures. Figures 14 and 15 illustrate reach and frequency when GRP is allocated solely to daypart 7. Its reach is less than half in comparison with the base case while its frequency is more than twice.

The resulting total category sales, Ocean Spray sales and share are shown in Figure 16, 17, and 18 respectively. They suggest that a point of GRP is more effective for early evening (4:30pm~7pm) and late night (10pm~1am) on weekdays and afternoon (1pm~4pm) on weekend for the twenty households sampled.

The study leads to an investigation of daypart allocation by spreading the weekly GRP among dayparts. A scheme proposed here chooses three dayparts which achieve the highest brand sales based on Figure 17 and re-distributes the total weekly GRP to the three according to their sales. They are 4:30pm~7pm and 10pm~1am on weekdays and 1pm~4pm on weekend, which are considered to generate high reach by covering different time frames. As seen in Table 4, the plan improves category and Ocean Spray sales by approximately 2.0% and 2.6% respectively.

Figure 12: Ocean Spray share for no ad, base, and twice the base cases





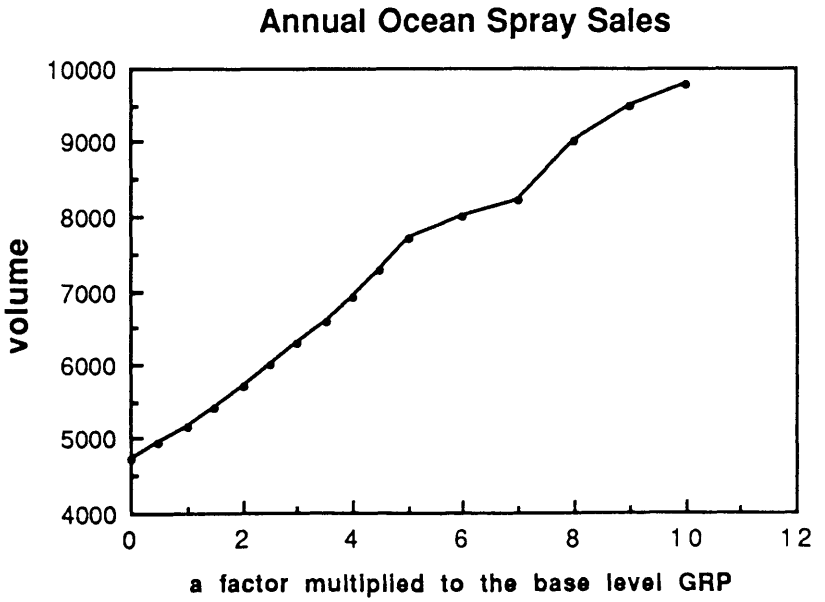


Figure 13: Annual sales of ad scenarios with GRP from 0 to 10 times the base level

Figure 14: Weekly reach when GRP is allocated solely to daypart 7

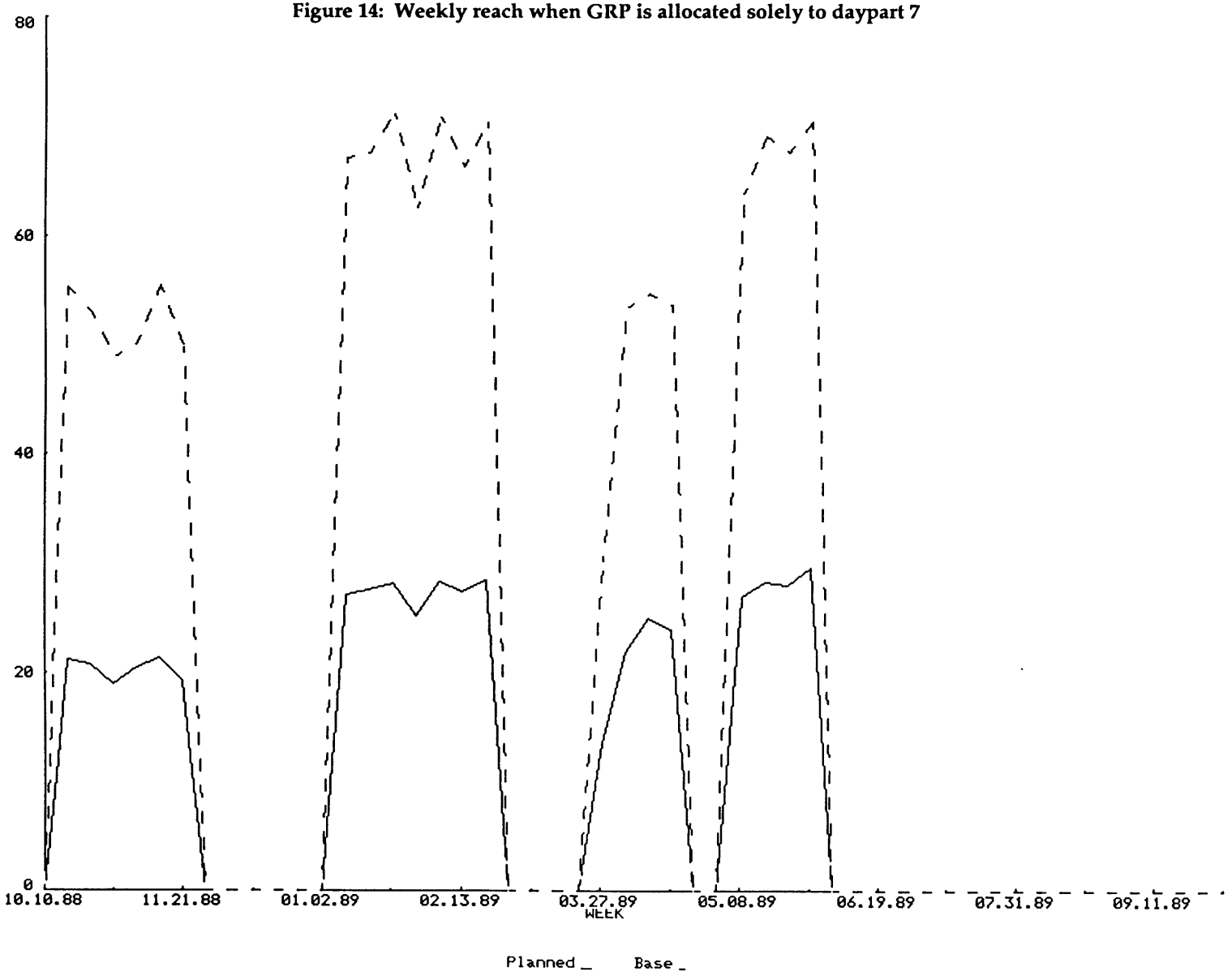


Figure 15: Weekly frequency when GRP is allocated solely to daypart 7

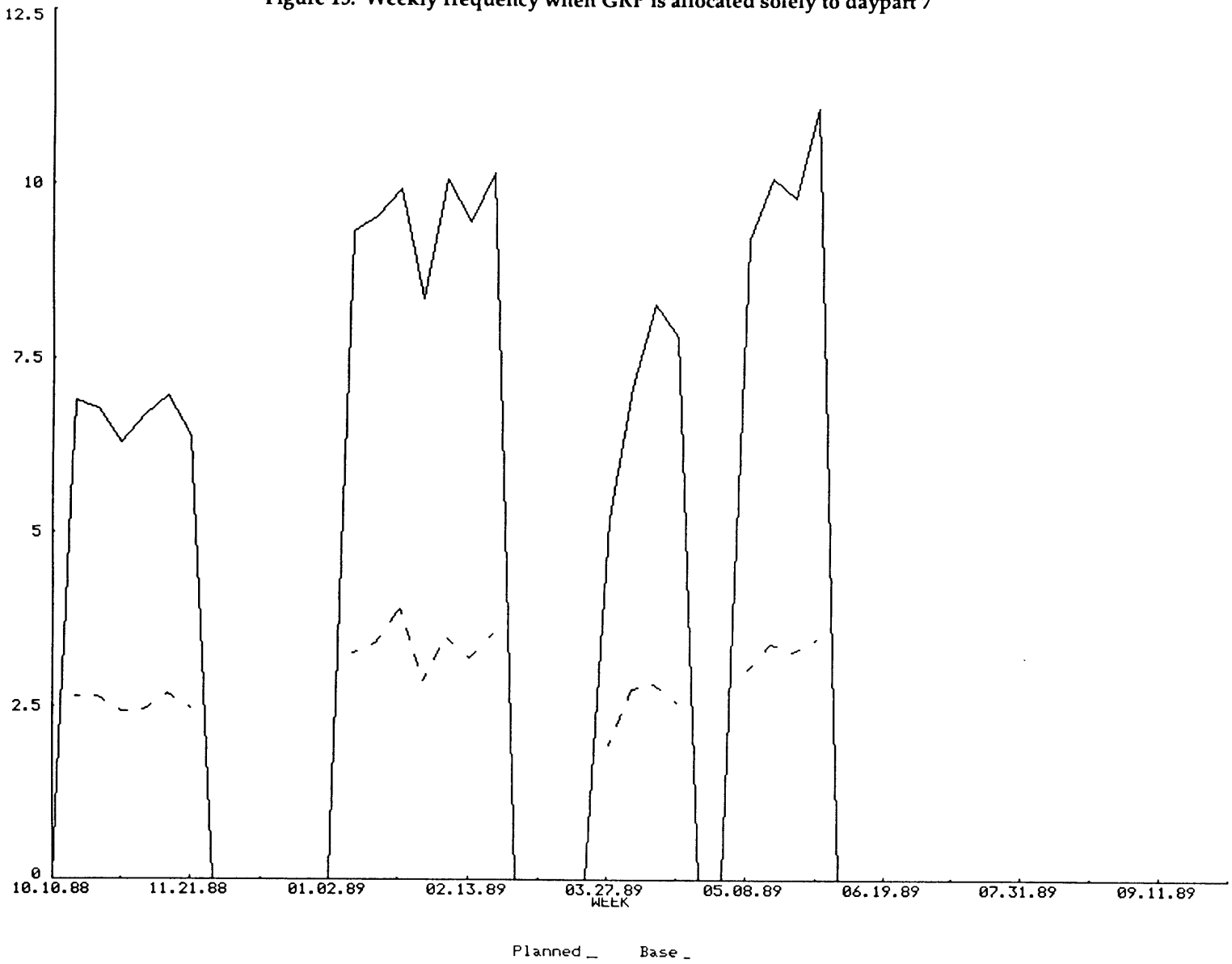


Figure 16: Total category sales when GRP is allocated to a single daypart

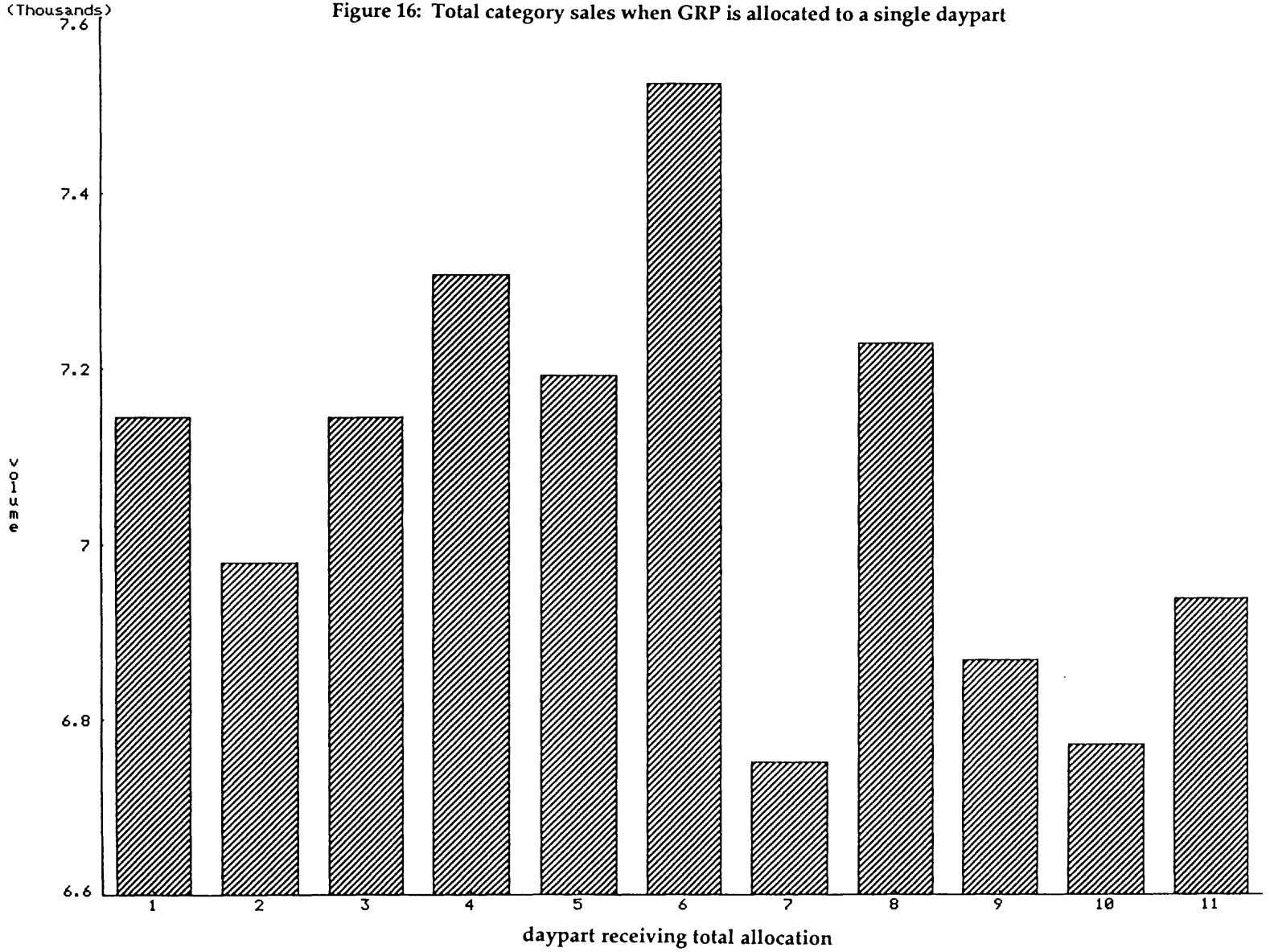


Figure 17: Total Ocean Spray sales when GRP is allocated to a single daypart

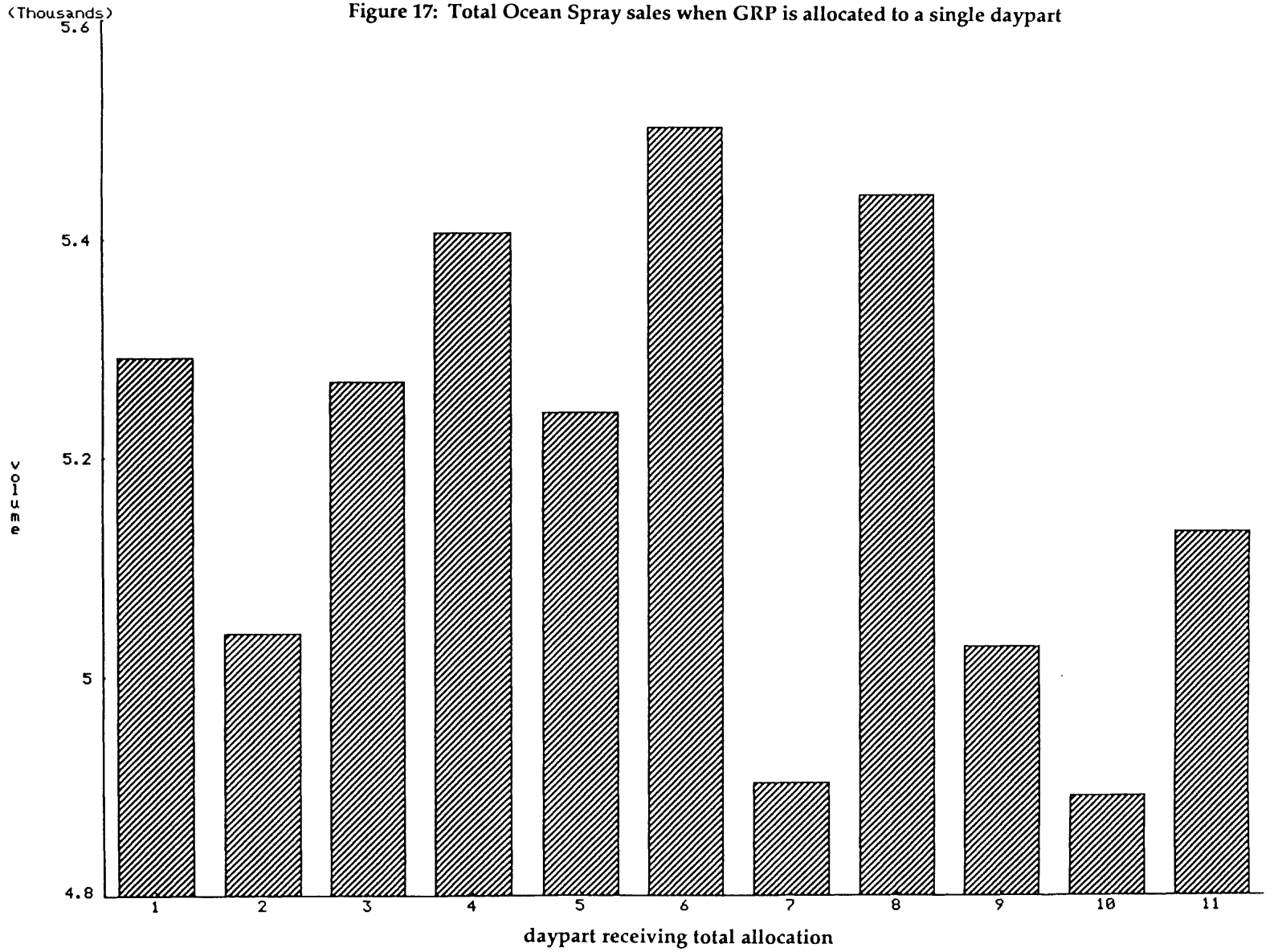


Figure 18: Ocean Spray share when GRP is allocated to a single daypart

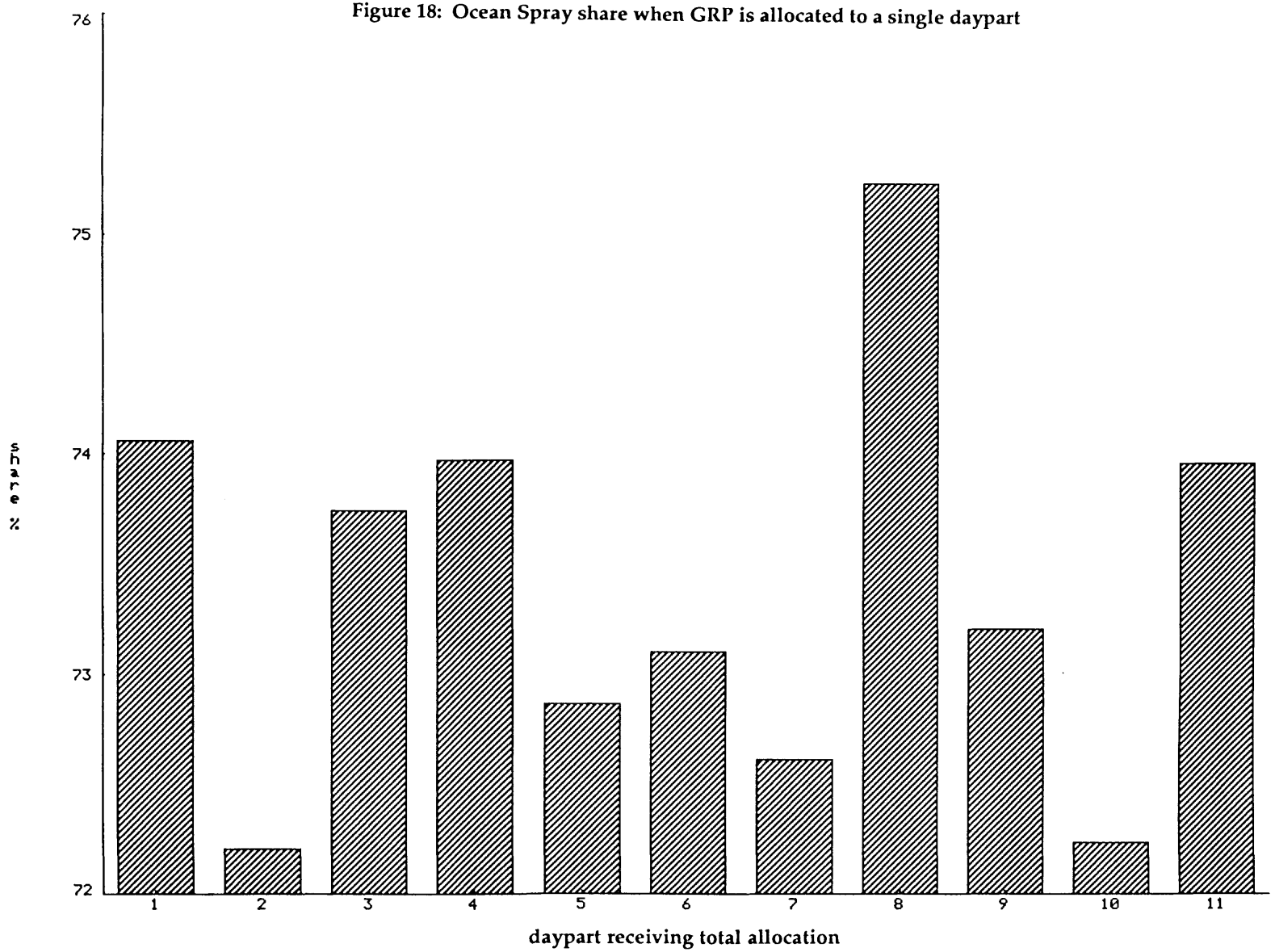


Table 4: Aggregate summary of the two daypart allocation schemes

measure	Base	Re-allocation
Category sales [oz]	7,087.7	7,231.6
(% change from base)		(+2.0%)
OS sales [oz]	5,177.3	5,311.5
(% change from base)		(+2.6%)
OS share [%]	73.05	73.45
OS vol / buyer [oz]	258.9	265.6
No. of purchase / buyer	7.31	7.44
(% change from base)		(+1.8%)
OS vol / purchase [oz]	35.4	35.7
(% change from base)		(+0.8%)

4.3 Advertising Pulsing

There has been much debate on whether pulsing of media advertising is more effective than constant spending over time. (Ackoff & Emshoff 1975, Rao & Miller 1975, Simon 1982, Mahajan & Muller 1986, Feinberg 1988) The issue is also directly related to the S-shaped aggregate response curve discussed earlier. Here, pulsing of various cycles (with a duty cycle[27] of 50%) have been implemented while keeping the same total GRP and its relative daypart allocation as the observed base case.

Table 5 is the aggregate results of the scenarios whose cycles are 52 weeks (26), 26 weeks (13), 13 weeks (6), 6-7 weeks (3), 2 weeks (1), and a constant level for 52 weeks, where the numbers in the parentheses indicate the duration of the positive constant pulse in weeks within the cycle. Hence, "6-7 weeks (3)" above means that 3 weeks of an ad flight is followed by no ad for 3 weeks, then another 3 weeks of a flight is followed by 4 weeks of no ad period, and this 13-week process is repeated four times over the year. Because the results of Table 5 are specific to our hypothesized advertising coefficients and are also confounded with other marketing mix variables, one must be careful in generalizing to other situations. Under the circumstance of the

[27] The duty cycle refers to a ratio of the duration of the high pulse to that of the whole period in a square wave.

current database, the pulsing of 6-7-week cycle --- 3 weeks of a flight followed by either 3 or 4 weeks of no ad --- produces the highest category as well as brand sales.

Table 5: Aggregate summary of the advertising pulsing scenarios

measure	52-week	26-week	13-week	6-7-week	2-week	constant
Category sales [oz]	7,072.6	7,076.2	7,125.5	7,149.0	7,140.7	7,140.8
OS sales [oz]	5,166.2	5,177.7	5,208.6	5,223.6	5,214.7	5,213.6
OS share [%]	73.04	73.17	73.10	73.07	73.03	73.01
OS vol / buyer [oz]	258.3	258.9	260.4	261.2	260.7	260.7
No. of purchase / buyer	7.30	7.30	7.35	7.37	7.36	7.36
OS vol / purchase [oz]	35.39	35.46	35.42	35.43	35.41	35.40

5. CONCLUSION

What have we learned from the study? Little stated as follows in his 1979 article.

> *"Looking ahead, new developments in measurement offer the possibility of resolving some of the outstanding modeling issues. ... The coupling of individual purchase information with observations of media exposure should permit ongoing response measurements ... Individual level measurements also seem required to examine hypotheses being generated from behavioral science. At the same time, the measurements must be tied into models.... In the next 5 to 10 years there will be abundant opportunities for understanding advertising processes better and putting this knowledge to work in improving marketing productivity."*

The advances in information technology in the past decade have brought us to the point where such sizable measurements are now readily available on a hard disk in a desktop PC. The time has come for more disaggregate analyses. The current research is an attempt at dissaggregate modeling for media planning by taking an advantage of the new data availability.

Input in this study is GRP by week and daypart, while outputs are brand sales, share, volume per buyer, number of purchases per buyer, and volume per purchase as well as exposure measures, weekly reach and frequency. The aggregate outputs are computed from category purchase and brandsize choice probability of each household at each purchase opportunity. The model has two parts. The first one is a household marketing mix model based on nested logit described in Part III, and the second is a Poisson ad exposure model calibrated on household media habits. Different advertising scenarios are simulated with the model by changing the GRP level, re-allocating GRP among dayparts, and re-scheduling flights. Computational efficiency is sought by introducing a new Taylor series simulation instead of a usual Monte Carlo simulation which requires multiple runs for stability.

We must realize that TV media planning covers just one part of the whole picture of a marketing mix model presented in Figure 1. The current model can be readily accommodated to simulate various store level marketing activities such as promotions and pricing on a weekly basis. However, there still exists numerous opportunities remaining.

REFERENCES

Aaker, David A. (1975)
"ADMOD: An Advertising Decision Model"
Journal of Marketing Research, vol. 12, February, 37-45

Abe, Makoto (1991)
"A Moving Ellipsoid Method for Nonparametric Regression and its Application to Logit Diagnostics using Scanner Data"
Forthcoming, Journal of Marketing Research, August 1991

Abraham, Magid M. and Leonard M. Lodish (1987)
"PROMOTER: An Automated Promotion Evaluation System"
Marketing Science, vol. 6, no. 2, Spring, 101-123

Ackoff, R. L. and J. R. Emshoff (1975)
"Advertising Research at Anheuser-Busch, Inc. (1963-68)"
Sloan Management Review, vol. 16, 1-16

Amemiya, Takeshi (1985)
"Advanced Econometrics"
Harvard University Press, Cambridge

Bass, Frank M. and R. T. Lonsdale (1966)
"An Exploration of Linear Programming in Media Selection"
Journal of Marketing Research, vol. 3, 179-188

Ben-Akiva, Moshe and Steve R. Lerman (1985)
"Discrete Choice Analysis"
MIT Press, Cambridge, MA

Blattberg, Robert C., and Alan Levin (1987)
"Modelling the Effectiveness and Profitability of Trade Promotions"
Marketing Science, vol. 6, no. 2, Spring, 124-146

Breiman, Leo and Friedman, Jerome H. (1985)
"Estimating Optimal Transformations for Multiple Regression and Correlation"
Journal of the American Statistical Association, vol. 80, no. 391, September, 580-619

Buja, Andreas, Trevor Hastie, and Robert Tibshirani (1989)
"Linear Smoothers and Additive Models"
The Annals of Statistics, Vol.17, No.2, 453-555

Bumbaca, Rico (1988)
"A Model of Discrete Choice Using Probability Density Estimation"
Master thesis, MIT Sloan School of Management, February 1988

Cacioppo, John T. and Richard E. Petty (1985)
"Central and Peripheral Routes to Persuasion: The Role of Message Repetition"
in Psychological Processes and Advertising Effects: Theory, Research, and Applications, L. F. Alwitt and A. A. Mitchell, eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 91-111

Carpenter, Gregory S., and Donald R. Lehmann (1985)
"A Model of Marketing Mix, Brand Switching, and Competition"
Journal of Marketing Research, vol. 22, no. 3, August, 318-329

Charnes, Cooper, Lerner, DeVoe, and Reinecke (1968)
"A Goal Programming Model for Media Planning"
Management Science, vol. 14, April, 423-430

Clarke, Darral G. (1976)
"Econometric Measurement of the Duration of Advertising Effect on Sales"
Journal of Marketing Research, vol. 13, November, 345-357

Cosslett, Stephen R. (1983)
"Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model"
Econometrika, Vol. 51, 765-782

Craig, C. Samuel, Brian Sternthal, and Clark Leavitt (1976)
"Advertising Wearout: An Experimental Analysis"
Journal of Marketing Research, vol. 13, November, 365-372

Devroye, L. and Gyorfi, L. (1985)
"Nonparametric Density Estimation: The L_1 View"
Wiley, NY

Donthu, Naveen and Rust, Ronald T. (1989)
"Estimating Geographic Customer Densities Using Kernel Density Estimation"
Marketing Science, Vol. 8, No. 2, Spring 1989, 191-203

Duncan, Gregory M. (1986)
"A Semi-Parametric Censored Regression Estimator"
Journal of Econometrics, 32, 5-34

Efron, Bradley (1981)
"Nonparametric Estimates of Standard Error: The Jackknife, the bootstrap and other methods"
Biometrika, Vol. 68, No. 3, 589-599

Efron, Bradley, and Gail Gong (1983)
"A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation"
The American Statician, Feb 83, Vol. 37, No. 1

Ehrenberg, A. S. C. (1972)
"Repeat Buying"
Amsterdam: North-Holland

Fader, Peter and James Lattin (1990)
"Accounting for Non-Stationary Household Preferences in a Cross-Sectional Model of Consumer Purchase Behavior"
Working Paper, Wharton School, University of Pennsylvania

Fader, Peter, James Lattin, and John D. C. Little (1990)
"Estimating Nonlinear Parameters in the Multinomial Logit Model"
MIT Marketing Center Working Paper 90-8

Feinberg, Fred M. (1988)
"Pulsing Policies for Aggregate Advertising Models"
MIT Doctoral Thesis, Sloan School of Management

Friedman, J. H. and W. Stuetzle (1981)
"Projection Pursuit Regression"
Journal of American Statistical Association, vol. 76, 817-823

Gensch, Dennis H. (1969)
"A Computer Simulation Model for Selecting Advertising Schedules"
Journal of Marketing Research, vol. 6, May, 203-214

Good, I. J. and Gaskins, R. A. (1980)
"Density Estimation and Bump-hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data"
Journal of American Statistical Association, Vol.75, 42-73

Guadagni, Peter M., and John D. C. Little (1983)
"A Logit Model of Brand Choice Calibrated on Scanner Data"
Marketing Science, vol. 2, no. 3, Summer, 203-238

Guadagni, Peter M., and John D. C. Little (1987)
"When and What to Buy: A Nested Logit Model of Coffee Purchase"
Working Paper WP 1919-87, Sloan School of Management, August

Gupta, Sunil (1988)
"Impact of Sales Promotions on When, What, and How Much to Buy"
Journal of Marketing Research, vol. 25, no. 4, November, 342-355

Gupta, Sunil (1991)
"Stochastic Models of Interpurchase Time With Time-Dependent Covariates"
Journal of Marketing Research, vol. 28, February, 1-15

Gurumurthy, K., and John D. C. Little (1989)
"A Price Response Model Developed from Perceptual Theories"
Working Paper, MIT Marketing Center WP89-5

Han, A. K. (1987)
"Nonparametric Analysis of a generalized Regression Model: The Maximum Rank Correlation Estimation"
Journal of Econometrics, vol. 35, 303-316

Hastie, Trevor and Robert Tibshirani (1986)
"Generalized Additive Models"
Statistical Science, vol. 1, no. 3, 297-318

Hastie, Trevor and Robert Tibshirani (1987)
"Generalized Additive Models: Some Applications"
Journal of the American Statistical Association, June , vol. 82, no. 398, 371-386

Hastie, Trevor and Robert Tibshirani (1990)
"Generalized Additive Models"
Chapman and Hall, New York

Hauser, John R. (1978)
"Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information Theoretic Approach"
Operations Research, vol. 26, May, 406-421

Hauser, John R., and Kenneth J. Wisniewski (1982)
"Application, Predictive Test, and Strategy Implications for a Dynamic Model of Consumer Response"
Marketing Science, vol. 1, no. 2, Spring, 143-180

Hausman, Jerry and Daniel McFadden (1984)
"A Specification Test for the Multinomial logit Model"
Econometrica, vol. 52, no. 5, September, 1219-1240

Jeuland, Abel P. (1978)
"A Multialternative-Multiattribute Choice Model with Individual Taste Differences"
Working Paper, Graduate School of Business, University of Chicago

Jeuland, Abel P., Frank M. Bass, and Gordon P. Wright (1980)
"A Multibrand Stochastic Model Compounding Heterogeneous Erlang Timing and Multinomial Choice Processes"
Operations Research, vol. 28, no. 2, March-April

Jones, J. Morgan, and Fred S. Zufryden (1980)
"Adding Explanatory Variables to a Consumer Purchase Behavior Model: An Exploratory Study"
Journal of Marketing Research, vol.17, no. 3, August, 323-334

Kahn, Barbara E., Donald G. Morrison, and Gordon P. Wright (1986)
"Aggregating Individual Purchases to the Household Level"
Marketing Science, vol. 5, no. 3, Summer, 260-268

Kalwani, Manohar U., Chi Kin Yim, Heikki J. Rinne, and Yoshi Sugita (1989)
"A Price Expectations Model of Customer Brand Choice"
Working Paper, Krannert Graduate School of Management, Purdue University

Kanetkar, Vinay, Charles B. Weinberg, and Doyle L. Weiss (1989)
"Price Sensitivity and Television Advertising Exposures: some Empirical Findings"
Working Paper, Faculty of Management, University of Toronto

Klein, R. W. and R. H. Spady (1988)
"Semiparametric Estimation of Discrete Choice Models"
Working Paper, Bell Communication Research

Krishnamurthi, Lakshman, and S. P. Raj (1985)
"The Effect of Advertising on Consumer Price Sensitivity"
Journal of Marketing Research, vol. 22, no. 2, May, 119-129

Krishnamurthi, Lakshman, and S. P. Raj (1988)
"A Model of Brand Choice and Purchase Quantity Price Sensitivities"
Marketing Science, vol. 7, no. 1, Winter, 1-20

Kuehn, Alfred A. (1962)
"Consumer Brand Choice - A Learning Process?"
Journal of Advertising Research, vol. 2, 10-17

Landwehr, James M, Daryl Pregibon, and Anne C. Shoemaker
"Graphical Methods for Assessing Logistic Regression Models"
Journal of American Statistical Association, Vol.79, No..385, 61-71

Lattin, James M (1987)
"A Model of Balanced Choice"
Marketing Science, vol. 6, no. 1, Winter, 48-65

Lattin, James M., and Randolph E. Bucklin (1989)
"Reference Effects of Price and Promotion on Brand Choice Behavior"
Journal of Marketing Research, vol. 26, no. 3, August, 299-310

Lilien, Gary L. (1974)
"A Modified Linear Learning Model of Buyer Behavior"
Management Science, vol. 20, no. 3, March, 279-285

Little, John D. C. and Leonard M. Lodish (1966)
"A Media Selection Model and Its Optimization by Dynamic Programming"
Industrial Management Review, vol. 8, 15-24

Little, John D. C. and Leonard M. Lodish (1969)
"A Media Planning Calculus"
Operations Research, vol. 17, no. 1, 1-35

Little, John D. C. (1975)
"BRANDAID: A Marketing-Mix Model. Part 1: Structure: Part II: Implementation"
Operations Research, vol.23, no. 4, 628-673

Little, John D. C. (1979)
"Aggregate Advertising Models: The State of the Art"
Operations Research, vol. 27, No. 4, 629-667

Lodish, Leonard M. (1971)
"Empirical Studies on Individual Response to Exposure Patterns"
Journal of Marketing Research, vol. 8, May, 212-218

Maddala, G. S. (1983)
"Limited-Dependent and Qualitative Variables in Econometrics"
Cambridge University Press, New York

Mahajan, V. and E. Muller (1986)
"Advertising Pulsing Policies for Generating Awareness for New Products"
Marketing Science, vol. 5, no. 2, 89-106

Mallows, C. L. (1973)
"Some Comments on Cp"
Technometrics vol. 15, 661-675

Manski, Charles F. (1975)
"Maximum Score Estimation of the Stochastic Utility Model of Choice"
Journal of Econometrics, 3, 205-228

Manski, Charles F. (1986)
"Semiparametric Analysis of Binary Response from Response-Based Samples"
Journal of Econometrics, 31, 31-40

Manski, Charles F. (1989)
"Estimation of Best Predictors of Binary Response"
Journal of Econometrics, 40, 97-123

Marron, J. S.. (1989)
"Automatic Smoothing Parameter Selection: A Survey"
in Semiparametric and Nonparametric Econometrics, ed. A. Ullah, Physica-Verlag Heidelberg

Matzkin, Rosa L. (1989)
"Estimation of Multinomial Models Using Weak Monotonicity Assumptions"
Working Paper, Cowles Foundation for Research in Economics, Yale University

McCullagh, P. and J. A. Nelder (1989)
"Generalized Linear Models"
2nd Edition, Chapman and Hall, New York

McFadden, Daniel, W. Tye, and K. Train (1976)
"An Application of Diagnostic Tests for the Independence of Irrelevant Alternatives Property of the Multinomial Logit Model"
Transportation Research Board Record, 637, 39-45

McFadden, Daniel (1981)
"Economic Models of Probabilistic Choice"
in Structural Analysis of Discrete Data with Econometric Applications
ed. C. F. Manski and D. McFadden, MIT Press, Cambridge

McFadden, Daniel (1985)
"Regression Based Specification Tests for the Multinomial Logit Model"
Journal of Econometrics, vol. 34, 63-82

Meyers, Raymond H. (1986)
"Classical and Modern Regression with Applications"
Duxbury Press, Boston, MA

Mosteller, Frederick and John W. Tukey (1977)
"Data Analysis and Regression"
Addison-Wesley; 1977

Nadaraya, E. A. (1970)
"Remarks on Non-parametric Estimates for Density Functions and Regression Curves"
Theory of Probability and Its Applications, Vol. 15, No. 1, 134-136

Nelder, J. A. and R. W. M. Wedderburn (1972)
"Generalized Linear Models"
Journal of Royal Statistical Society A, vol. 135, 370-384

Neslin, Scott A., Caroline Henderson, and John Quelch (1985)
"Consumer Promotions and the Acceleration of Product Purchases"
Marketing Science, vol. 4, no. 2, Spring, 147-165

Neslin, Scott A., and Robert W. Shoemaker (1989)
"An Alternative Explanation for Lower Repeat Rates After Promotion Purchases"
Journal of Marketing Research, vol. 26, no. 2, May, 205-213

Parzen, E. (1962)
"On Estimation of a Probability Density Function and Mode"
Annals of Mathematical Statistics, Vol.33, 1065-1076

Pedrick, James H., and Fred S. Zufryden (1990)
"Evaluating the Impact of Advertising Media Plans: A Model of Consumer Purchase Dynamics Using Single-Source Data"
Working Paper, Department of Marketing, University of Southern California

Prakasa Rao, B. L. S. (1983)
"Nonparametric Functional Estimation"
Academic Press, New York

Rao, A. G. and P. B. Miller (1975)
"Advertising/Sales Response Functions"
Journal of Advertising Research, vol. 15, 7-15

Rust, Roland T. (1985)
"Selecting Network Television Advertising Schedules"
Journal of Business Research, vol. 13, 483-494

Rust, Roland T. (1988)
"Flexible Regression"
Journal of Marketing Research, vol. 25, February, 10-24

Sawyer, Alan (1981)
"Repetition, Cognitive Responses and Persuasion" *in Cognitive Responses in Persuasion*
R. E. Petty, T. M. Ostrom, and T. C. Brock, eds. Hillsdale, NJ: Lawrence Erlbaum Associates 237-261

Silverman, Bernard W. (1986)
"Density Estimation for Statistics and Data Analysis"
Monographs on Statistics and Applied Probability, Chapman and Hall, New York

Simon, J. L. (1969)
"New Evidence for No Effect of Scale in Advertising"
Journal of Advertising research, vol.9 38-41

Simon, H. (1982)
"ADPULS: An Advertising Model with Wearout and Pulsation"
Journal of Marketing Research, vol. 19, 352-363

Stoker, Thomas M. (1986)
"Consistent Estimation of Scaled Coefficients"
Econometrica, Vol. 54, No. 6, November, 1461-1481

Stone, Charles J. (1985)
"Additive Regression and Other Nonparametric Models"
The Annals of Statistics, vol. 13, no. 2, 689-705

Tellis, Gerard J. (1988)
"Advertising Exposure, Loyalty, and Brand Purchase: A Two-Stage Model of Choice"
Journal of Marketing Research, vol. 25, no. 2, May, 134-144

Ullah, Aman (1988)
"Non-parametric Estimation of Econometric Functionals"
Canadian Economics Association, Vol. 21, No. 3, 625-658

Wagner, Udo, and Alfred Taudes (1986)
"A Multivariate Polya Model of Brand Choice and Purchase Incidence"
Marketing Science, vol. 5, no. 3, Summer, 219-244

Watson, Geoffrey S. (1964)
"Smooth Regression Analysis"
Sankhya A, Vol. 26, No.4, 359-372

Wegman, E. J. and Wright, I. W. (1983)
"Splines in Statistics"
Journal of American Statistical Association, Vol.78, 351-365

Wheat, Rita D., and Donald G. Morrison (1990)
"Assessing Purchase Timing Models: Whether or Not is Preferable to When"
Marketing Science, vol. 9, no. 2, Spring, 162-170

Winer, Russell S. (1986)
"A Reference Price Model of Brand Choice for Frequently Purchased Products"
Journal of Consumer Research, vol. 13, September, 250-256

Zangwill, W. I. (1965)
"Media Selection by Decision Programming"
Journal of Advertising Research, vol. 5, September, 30-36

Zufryden, Fred S. (1987)
"A Model for Relating Advertising Media Exposures to Purchase Incidence Behavior Patterns"
Management Science, vol. 33, no. 10, October, 1253-1266