Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data:

Technical Report

by

Alan L. Montgomery
The Wharton School
University of Pennsylvania
Marketing Department
1400 Steinberg Hall-Dietrich Hall
Philadelphia, PA 19104-6371

e-mail: alm@wharton.upenn.edu

August 1995
Revised: June 1996
Second Revision: December 1996

The author would like to thank Peter Rossi and Steve Hoch for their valuable input. He also thanks Mark Bergen, Pete Fader, Kris Helsen, Rob McCulloch, Jagmohan Raju, and George Tiao for their comments, Dominick's Finer Foods, Information Resources Inc., and Market Metrics for their assistance and provision of data, and Xavier Dreze and Mary Purk for their indispensable help throughout. Financial support for this work was provided by the Micro-Marketing Project at the Graduate School of Business, University of Chicago and the Wharton School of the University of Pennsylvania.

Copyright © 1996 by Alan Montgomery, All rights reserved

Appendix A: Description of Dataset

Our dataset was collected as part of the Micro-Marketing Project at the University of Chicago (Dreze, Hoch, and Purk 1993). It is composed of data from three sources.

Store-level Scanner Data: DFF provided weekly UPC-level scanner data for all 88 stores in the chain for up to three years. The scanner data includes unit sales, retail price, profit margin, and a deal-code. Out of these 88 stores, five have limited historical data, so we concentrate on the remaining 83 stores. To verify the correctness of the data, comparisons across stores and categories were made for each week. Certain weeks in which the integrity of the data was in doubt were removed from the sample. Also one brand was introduced in the early part of the data (Florida Gold), and another brand (Citrus Hill) is removed in the later part. Consequently we consider the middle 121 weeks of the sample period (June 1990 through October 1992), to avoid introduction and withdrawal effects.

There are 33 UPCs in the category. In order to a create a more manageable number of products we create eleven aggregates from the original UPC level data that have similar pricing and promotional strategies.[a] The UPCs within a product aggregate differ only by flavoring, additives, or packaging (eg., regular, pulp, or calcium). The price of the aggregate is computed as a price index (i.e., an average weighted by market share) over all the UPCs that comprise the aggregate. The movement of the aggregate is computed as the sum of the movement (standardized to ounces). Prices within each aggregate are approximately proportional, therefore little pricing information is lost. Moreover we can still speak about profit maximization since we assume the relative prices of the items within an aggregate are fixed.

There is a natural division of products into three price-quality tiers: the premium brands (made from freshly squeezed oranges), the national brands (reconstituted from frozen orange juice concentrate), and the store brands (Dominick's private label). There is quite a bit of disparity in prices across the tiers, which leads to large differences in wholesale costs, even though the profit margins appear similar. An initial indication that store

a. One common technique is to consider only a specific size or the top 10 UPCs. The problem with this method is that we are ignoring a large percentage of the category, and the retailer must consider pricing the entire category and not simply a subset of the category. Our method is to consider those UPCs which correspond with the top 95% of category sales in the market, and then form aggregates of those groups of UPCs which have correlations of at least .9 with each of the other prices included in the aggregate.

differences are present is the variation of market shares across stores. Dominick's 64 ounce OJ brand has an average market share of 13.6%, but the market shares across stores range anywhere from a minimum of 5.6% to a maximum of 20.9%.

Promotional Data: Information about feature advertising in weekly newspaper fliers is provided by IRI's Infoscan, which provides an estimate of all commodity volume of a particular UPC that received feature advertisement. In-store promotion is measured using a deal code provided in DFF's store-level scanner database. The deal code is a dummy variable which shows whether there was a bonus-buy tag on the shelf or an in-store coupon. Since these promotional variables are at the UPC level, we create indices of the feature and deal variables for each aggregate similarly to that of price.

Store Trading Area Data (Competitive/Demographic Characteristics): Market Metrics, a leading firm in the use of demographic data, used block level data from the U.S. Census to compute a store's trading area. A store's trading area refers to a geographical area around the store. It is calculated by finding the number of people needed to sustain a given level of sales for this area. Geographical boundaries (such as roads, railroad tracks, rivers, etc.) are considered when this trading area is formed. The demographic composition of the store's trading area is computed by summing up the assigned proportion of each of the U.S. Census blocks within the prescribed trading area.

A total of eleven demographic and competitive variables are used to characterize a store's trading area. These variables summarize all major categories of information that are available. For a further discussion of variable selection refer to HKMR (1995). We have two measures of competition, distance (in miles) and relative volume, for two different types of competitors, warehouse (EDLP format) and supermarket (Hi-Lo format) stores. Distance is doubled in urban areas to reflect poorer driving conditions, which approximates Market Metrics measure of driving times. Relative volume is the ratio of sales in the competitor to that of the Dominick's store. The warehouse competitor variables are computed with respect to the nearest warehouse store, and the supermarket competitor variables use an average of the nearest five supermarket competitors.

Appendix B: Store-level Systems in a Hierarchical Model

To make the procedure as general as possible we rewrite our demand system in SUR form. We use a SUR model and not a simple multivariate regression since the feature and deal terms are different for each equation, although the price vector is identical. The system becomes:

$$\boldsymbol{y}_s = X_s \boldsymbol{\beta}_s + \boldsymbol{e}_s \quad , \quad e_s \sim N(\mathbf{0}, \Sigma_s \otimes I_T) \tag{B.1}$$

Here the *s* subscript denotes an individual store, and the dimension of the $\boldsymbol{y}$ vector is *M* brands by *T* weeks. In rewriting the model we have stacked the vector of observations for each brand:

$$\boldsymbol{y}_s = \begin{bmatrix} \boldsymbol{q}_{1s} \\ \boldsymbol{q}_{2s} \\ \vdots \\ \boldsymbol{q}_{Ms} \end{bmatrix}, \ \boldsymbol{q}_{is} = \begin{bmatrix} \ln(q_{i1s}) \\ \ln(q_{i2s}) \\ \vdots \\ \ln(q_{iTs}) \end{bmatrix}, \ X_s = \begin{bmatrix} X_{1s} & & & \\ & X_{2s} & & \\ & & \ddots & \\ & & & X_{Ms} \end{bmatrix}, \ X_{is} = \begin{bmatrix} 1 & p_{11s} & \cdots & p_{M1s} & f_{i1s} & d_{i1s} \\ 1 & p_{12s} & \cdots & p_{M2s} & f_{i2s} & d_{i2s} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 1 & p_{1Ts} & \cdots & p_{MTs} & f_{iTs} & d_{iTs} \end{bmatrix} \tag{B.2}$$

Note that $\boldsymbol{q}_{is}$ in equation (B.2) refers to the vector of log movement for a given brand over all weeks, whereas $\boldsymbol{q}_{ts}$ in equation (2.2) it refers to the vector of log movement across all brands for a given week.

The first stage of our hierarchical model is given by the system of demand equations in equation (2.1).

$$\Sigma_s^{-1} = G_{\Sigma_s} \sim \textit{Wishart}(\nu_\Sigma, \bar{V}_\Sigma^{-1}) \tag{B.3}$$

To complete this stage, we also specify natural conjugate priors on the error covariance matrix Σ_s:

The second stage refers to the hyper-distribution from which the parameters for each store are drawn and is defined in equation (2.5). To complete this second stage, we include a prior distribution on the covariance matrix of the second stage V_β:

$$V_\beta^{-1} = G_\beta \sim \textit{Wishart}(\nu_\beta, \bar{V}_\beta^{-1}) \tag{B.4}$$

The motivation for representing V_β with a prior distribution instead of specifying it directly is to allow for some

uncertainty in the amount of commonalities across stores.

The relationships between the demand parameters and the demographic and competitive variables are contained within the $Z_s\theta$ term in the second stage of our hierarchical model, which are stated in equation (2.6). Since all these relationships are linear, they can easily be incorporated into the Z_s matrix. We can partition Z_s and θ into constants and demographic components:

$$Z_s = [\, Z_c \; Z_{ds} \,] \quad , \quad \theta' = [\bar{\beta}' \;\; \gamma' \,] \tag{B.5}$$

Where the vector of chain-wide averages in the hyper-distribution is:

$$\bar{\beta}' = [\, \bar{\alpha}' \; vec(\bar{H})' \; \bar{\xi}' \; \bar{\Psi}' \; \bar{\phi}' \,] \tag{B.6}$$

and the relationships with the demographic and competitive variables are given by:

$$\gamma' = [\, \gamma_a' \; \gamma_o' \; \gamma_c' \; \gamma_f' \; \gamma_d' \,] \tag{B.7}$$

The Z_c matrix is composed of 1's and 0's and represents the constant vectors and therefore is the same for each store. In our model we let each β_i have its own intercept, hence Z_c is the identity matrix with order 154. If certain elements are to be ``shrunk" toward one another then the corresponding elements in a particular column are both set to 1, and the other elements set to 0.

Since the demographic data vector for each store is the same for all the parameters, the construction of the Z_{ds} matrix can be simplified using the following relationship:

$$Z_{ds} = d_s' \otimes Z_k \tag{B.8}$$

The Z_k matrix is constructed in an analogous manner to Z_c, except that it summarizes the systematic relationships. In our analysis the Z_k matrix has 5 columns. To illustrate this matrix consider the column which corresponds to the own-price sensitivities, if the parameter is an own-price sensitivity then the element is set to 1, otherwise the element is 0. Geometrically this allows for every coefficient to have its own intercept, but there is a common slope for the own-price elasticities inside each quality tier.

The third stage of our model expresses the prior on the hyper-parameters:

$$\theta \sim N(W\bar{\theta} , V_\theta) \tag{B.9}$$

In our specification we will employ a diffuse third stage. But an informative prior on θ would specify prior beliefs about chain-wide tendencies or demographic and competitive effects on parameter variation. The W matrix is included to make the specification of this prior more flexible.

B.1 Specification of Priors

The analyst must supply the following parameters and data:

$$X_s , Y_s , Z_s , W \quad ; \quad \nu_\Sigma , \bar{V}_\Sigma^{-1} , \nu_\beta , \bar{V}_\beta^{-1} , \bar{\theta} , V_\theta \tag{B.10}$$

Our priors on the error covariance matrices, Σ_s, and mean of the hyper-distribution, θ, are chosen to be are diffuse relative to the sample. The mean and degrees of freedom for the prior on Σ_s are:

$$\bar{V}_\Sigma = .1\, I \quad , \quad \nu_\Sigma = 1 \tag{B.11}$$

The parameters of the prior on our hyper-distribution θ are:

$$W = I \quad , \quad \bar{\theta} = 0 \quad , \quad V_\theta = 10^6 I \tag{B.12}$$

The most crucial prior will be on V_β, which reflects the strength of the commonalities across the stores. Notice that in our case the number of stores is less than the dimension of V_β. Therefore to form a proper posterior distribution, we will need to have an informative prior. Although as more information is added (more weeks of observations), the individual estimates of the β's will dominate this prior in deriving the posterior distribution of θ and V_β.

The motivation of the parameterization of our prior on V_β is to shrink our parameter estimates somewhere between the pooled and individual LS estimates. We set $\bar{V}_\beta$ to a diagonal matrix. To allow for proper scaling of the different coefficients we set the diagonal elements equal to the product of the variance least squares estimates from the individual store models, s_i^2, and a scaling parameter, k_i:

$$\bar{V}_\beta = \nu_\beta \bar{V}^*_\beta = \nu_\beta \begin{bmatrix} k_1^2 s_1^2 & & & \\ & k_2^2 s_2^2 & & \\ & & \ddots & \\ & & & k_p^2 s_p^2 \end{bmatrix} \tag{B.13}$$

where p is the dimension of β_s (or 154, i.e., 11 equations with 14 parameters each). We also set ν=160. The relationship between V_β^{-1} and $\bar{V}_\beta^{-1}$ can be seen by examining the mean and covariance matrix of this prior distribution: $E[V_\beta^{-1} \mid \nu_\beta, \bar{V}_\beta^{-1}] = \nu_\beta \bar{V}_\beta^{-1} = \bar{V}^*_\beta$ and $Var[V_\beta^{-1}] = 2/\nu_\beta \bar{V}^{*-1}_\beta \otimes \bar{V}^{*-1}_\beta$. For example, if k_i=.1 then our prior states that the expected standard deviation of store-specific random variation of our estimates will be 10% of those of the LS estimates. (Note that parameter variation is also be induced by the demographics.)

We are primarily concerned with parameter variation across stores. We do not try to shrink the parameters within a store (i.e., across products) closer to each other. Therefore, we let each store and brand have their own intercepts. Also to avoid a great deal of shrinkage in the constants we set the scaling parameters of these parameters to one or k, whichever is greater. For simplicity we set the scaling parameters for the other parameters equal, k_i=max(1,k) for constants and k_i=k for all other parameters. We choose a value of k=.1 to reflect our prior beliefs that more shrinkage is helpful and reflect the fact that pooled models tend to perform well compared with individual store models.

Appendix C: Estimating Hierarchical Bayes Models with the Gibbs Sampler

The nested structure of the model makes it easy to form the hierarchical model, although estimation presents another problem. Even with natural conjugate priors an exact solution of the posterior distribution is not known[b]. The difficulty in solving the marginal posteriors is due to the use of the Wishart distribution. This suggests estimation can be done using an normal approximation to the posterior by replacing Σ and V_β with point estimates (Montgomery 1994). An alternative estimator is to find a numerical solution. Unfortunately the high dimension of the integral makes it difficult to find a solution using conventional numerical integration techniques.

Since these approximations may be questionable and numerical integration is not feasible we make use of a new technique in computational statistics, known as the Gibbs sampler, to estimate the marginal posterior distributions. For a good introduction to the Gibbs sampler see Casella and George (1992). The Gibbs sampler requires randomly sampling from each of the conditional distributions sequentially. Due to the hierarchical structure of the model these conditional distributions can be readily computed. It has been shown by Gelfand and Smith (1990) and Gelfand et al. (1990) that these draws converge in distribution to the posterior marginal distributions. Although a disadvantage of the Gibbs Sampler is its intensive computer requirements.

Our implementation of the Gibbs Sampler involves the following three steps:

1. Select starting values for the parameters of the marginal posterior distributions. (We will use the least squares estimates of these parameters.)
2. Generate N_1+N_2 sets of random numbers with each set being drawn as:

$$\beta_s^{(k)} \veebar p(\beta_s \mid \Sigma_s^{(k-1)}, \theta^{(k-1)}, V_\beta^{(k-1)}) \quad for\ s = 1,\cdots,S \tag{C.1}$$

$$\Sigma_s^{(k)} \veebar p(\Sigma_s \mid \beta_s^{(k)}, \theta^{(k-1)}, V_\beta^{(k-1)}) \quad for\ s = 1,\cdots,S \tag{C.2}$$

b. The analytical solution which integrates Σ_s out of the joint distribution of β to derive the posterior distribution is not known. The difficulty arises as a result of the Wishart priors on Σ_s and V_β. To understand this problem, we refer the reader to the simpler case of trying to solve a single stage SUR model (Zellner 1971, pp. 240-6) for which the analytical solution is not known either.

$$\theta^{(k)} \leftarrow p(\theta \mid \beta_1^{(k)}, \cdots, \beta_S^{(k)}, \Sigma_1^{(k)}, \cdots, \Sigma_S^{(k)}, V_\beta^{(k-1)}) \tag{C.3}$$

$$V_\beta^{(k)} \leftarrow p(V_\beta \mid \beta_1^{(k)}, \cdots, \beta_S^{(k)}, \Sigma_1^{(k)}, \cdots, \Sigma_S^{(k)}, \theta^{(k)}) \tag{C.4}$$

Where $x \leftarrow p(x)$ denotes x as a draw or simulated value from the density $p(x)$, and k denotes the iteration. We set N_1=100 and N_2=2000, see Montgomery (1994) for a further discussion of convergence.

3. Use the last N_2 sets of draws to estimate the posterior marginal distributions. For example if we are interested in the posterior mean and variance, we could compute the sample mean and variance of the final N_2 draws and use this as our estimate. As long as the number of draws is large our estimation error will be small.

Step 2 requires that we solve the conditional distributions of each parameter. These solutions are readily available due to the model's hierarchical structure and the affine nature of the normal and Wishart distributions (Anderson 1984, pp. 84 and 268-269). Simulating draws from each of these distributions can be done using a standard statistical library. The solutions to the conditional distributions are:

1. β_s is a SUR model

$$\beta_s \mid \Sigma_s, \theta, V_\beta \sim N(H(X_s'(\Sigma_s^{-1} \otimes I_{T_s})y_s + V_\beta^{-1} Z_s \theta), H), \quad H = (X_s'(\Sigma_s^{-1} \otimes I_{T_s})X_s + V_\beta^{-1})^{-1} \tag{C.5}$$

2. Σ_s is drawn from an inverted Wishart distribution

$$\Sigma_s^{-1} \mid \beta_s, \theta, V_\beta \sim W(\nu_\Sigma + T_s, (\bar{V}_\Sigma + \hat{E}_s'\hat{E}_s)^{-1}), \quad \hat{E}_s[,i] = y_{is} - x_{is}'\beta_{is} \tag{C.6}$$

3. θ is a multivariate regression

$$\theta \mid \beta, \Sigma_s, V_\beta \sim N(H(Z'(I_S \otimes V_\beta^{-1})\beta + V_\theta^{-1}\bar{W}\bar{\theta}), H), \quad H = (Z'(I_S \otimes V_\beta^{-1})Z + V_\theta^{-1})^{-1} \tag{C.7}$$

4. V_β is drawn from an inverted Wishart distribution

$$V_\beta^{-1} \mid \beta, \Sigma, \theta \sim W(\nu_\beta + S, (\bar{V}_\beta + \Sigma(\beta_s - Z_s\theta)(\beta_s - Z_s\theta)')^{-1}) \tag{C.8}$$

These conditional distributions are understood to also depend upon the prior parameters and the data in (C.8), and

Z is a block diagonal matrix with $Z_1, Z_2, \ldots Z_S$ on the diagonal.

C.1 Estimating the Marginal Posterior Distribution of Expected Profits

An added benefit of the Gibbs draws is that we may compute an estimate of the marginal posterior distribution of the expected profit function which incorporates the uncertainty of the parameter estimates. Traditionally the posterior means of the parameter estimates are substituted into the profit function. Blattberg and George (1992) show that this method does not lead to an optimal pricing solution due to the nonlinearity of the profit function. Therefore, our procedure does not suffer from the drawbacks of the traditional method.

We save the last N_2 Gibbs iterations and evaluate our profit function at each of these iterations, which can be treated as draws from the marginal posterior profit function. If we wish to compute an estimate of the mean of this marginal posterior profit function, we can evaluate the profit function for each Gibbs iteration and divide by the number of iterations. Since these calculations are highly computer intensive when applying numerical optimization techniques, we use every 16th Gibbs iteration, for a total of 125 iterations. An estimate of total expected chain profits, $E[\Pi]$ from (3.1), is computed by evaluating (3.1) for each of the 125 iterations and then taking the average of these draws. The optimizations are computed by evaluating the value and derivates of this estimate. Additionally, to estimate $Pr(E[\pi_a]>E[\pi_b])$, where a and b denote different strategies, we evaluate $E[\pi_a]-E[\pi_b]$ for each Gibbs draw and compute the percentage of positive draws.