Statistics and Marketing

Peter E. Rossi and Greg M. Allenby

Statistical research in marketing is heavily influenced by the availability of different types of data. The last ten years have seen an explosion in the amount and variety of data available to market researchers. Demand data from scanning equipment has now become routinely available in the packaged goods industries. Data from e-commerce and direct marketing is growing at an exponential rate and provide coverage to a wide assortment of different products. Web-based technology has dramatically lowered the cost of survey research. Web-browsing data is an important new source of information about consumer tastes and preferences which is becoming available for a large fraction of the total consumer population. In this vignette, we explore some of the implications of this data explosion for the development of statistical methodology in marketing with primary emphasis on the explosion in demand data.

Scanning equipment has provided the market researcher with a national panel of stores in addition to panels of households, altering the focus of marketing research. This data has stimulated a large literature on applied demand and discrete choice modeling. Demand models at the store level typically take the form of multivariate regression models in which demand for a vector of products is related to marketing variables such as prices, displays and various forms of advertising. At the household level, demand is discrete and a wide variety of multinomial logit and probit models have been applied to the data.

Early experience with scanner data revealed that households have very different patterns of buying behavior that cannot be explained just by differences in the marketing environment. Some households, for example, exhibit strong brand loyalties while other households readily switch brands when prices are lowered. Even at the store level, large differences have been detected in price and local advertising sensitivity. Initial observations of store and consumer heterogeneity created considerable interest in models of observed and unobservable heterogeneity, primarily of the random effects form. The development and application of random effect models in marketing has been dictated in large degree by the available inference technology.

The first paper in this area by Kamarkura and Russell (1989) used a finite mixture model of heterogeneity in a logit framework. Kamarkura and Russell postulate a discrete

multivariate distribution for the intercepts and marketing mix variable coefficients in a multinomial logit model. The ease of computation of finite mixture models produced a stream of papers applied this idea in a wide variety of other model contexts. Until quite recently, it was not computationally practical to investigate multivariate continuous models of heterogeneity in a discrete choice setting. Simulation-based Maximum Likelihood (Geweke, 1989) and Markov Chain Monte Carlo (MCMC) Bayesian inference methods (Gelfand and Smith, 1990) have now made this possible. Recent work (Lenk, DeSarbo, Green and Young (1996) and Allenby and Rossi (1999)) has documented that finite mixtures provide poor approximations to the high dimensional distributions necessary to capture heterogeneity not only in the intercepts but also in the slopes of the multinomial choice models. Rossi, McCulloch and Allenby (1996) use a multivariate normal distribution of random effects in a Bayesian hierarchical setting. Observable household characteristics such as demographics are included in the model by allowing the mean of the random effects distribution to depend on these variables. These observable characteristics account for only a small amount of the total household heterogeneity. Brand preferences and marketing mix sensitivities are therefore revealed primarily by household purchase behavior.

A distinguishing characteristic of random-effects applications in marketing is the interest not only in the hyperparameters of the random-effects distribution but in making inferences about household parameters. Marketers can use household level parameters to target households for customized promotions or to cross-sell other products. Even at the store level, there is considerable interest in the allocation of marketing efforts across different geographic areas which requires the building of store or market level models.

The sampling-theoretic approaches to random effects models typically average the model likelihood over the random effects distribution and provide no natural way to make household level inferences. In contrast, Bayesian hierarchical models provide a natural setting in which inferences can be made concerning both the parameters of the random effects distribution and the household or store level parameters. The combination of ability to make inferences about the random parameter draws and computational tractability for even very high dimensional problems has made Bayesian hierarchical methods very popular in the empirical marketing literature (see Allenby and Lenk (1994), Montgomery (1997), Allenby, Arora and Ginter (1998), Ainslie and Rossi (1998), Boatwright and Rossi (1999),

Neelamegham and Chintagunta (1999), Allenby, Leone and Jen (1999), Bradlow and Schmittlein (1999)).

Just as the current statistical literature in marketing was stimulated by the availability of panel data, we expect the future course of research to be influenced by the growth in customer transaction databases. Direct marketers, credit card and financial service marketers, and retailers are now compiling large databases comprised of customer transactions and marketing contacts. These databases range from a complete record of all customer purchases along with information on other alternatives considered to much more incomplete data in which only the purchases are recorded (as in most frequent shopper or loyalty programs) or where only indirect evidence of product preferences is available (as in Internet Web-browsing data).

One of the most basic uses of customer transaction data is to find customers who are likely to express interest in or purchase a specific product. For example, a marketer could use credit usage data to find card holders who are likely to be interested in an upscale Italian restaurant and then target promotion activities at this subset. Another common example is the use of house file information in catalogue purchases to identify customers who are likely to respond to a catalogue mailing. The challenge is essentially a classification task. Can we construct a set of variables summarizing past transaction data and choose a model form which effectively locates the customers we want to reach? The industry uses a variety of standard regression and logit models for the most part. Given the extremely large set of possible variables and functional forms possible, many of the classification problems have a huge variable selection problem associated with them. Also, there is no particular theoretical reason to believe that standard linear regression or logistic regression models have the right functional form. Recent developments of the statistics literature on variable selection and CART models might find very fruitful application here (George and McCulloch (1995) and Chipman et al (1998)). It may also be useful for firms to build upon the current investment in regression models by considering regression models with coefficients which change depending on the location in the explanatory variable space. Neural networks models (c.f. Ripley 1996)) have already found application in predictive model building with marketing data.

In many instances, firms need to predict the response to a new or modified offering, ranging from new prices for old brands to entirely new offerings. A direct mail merchant

who wants to predict the effects of a price change, for example, needs information about prices of competing brands. While this information is often not available in a firm's transaction database, it can be imputed by augmenting the transaction data with data from other sources. For example, a firm could obtain the data for a subset of customers by recruiting a panel and having them record their purchases and prices for the other brands that were considered at the time. Once a joint model of all causal variables is available for a subset of customers, estimates of price sensitivity and brand preference can be generated from transaction data by constructing the conditional distribution of these variables given the available information.

Even when data are available to identify the parameters in a customer demand model, the resulting estimates are sometimes insufficient for marketing decisions. The brand intercepts, for example, reflect consumer preference for brands net the effects of price and the other variables that change through time. These preferences are driven by product attributes, the perceived performance of the brand on these attributes, and the importance that consumers attach to them. Furthermore, the importance of product attributes to a consumer can be traced to the concerns, interests and motives they bring to the problem for which the brand is relevant. Understanding these more primitive constructs is important in many marketing decisions, such as market segmentation, advertising and product development. A valuable area for future research is the merging of survey data with transaction data. For example, in choice models the brand intercepts reflect preference net the effect of price and other variable that change over time. Marketers want to relate these preferences to product attributes. For many products these attributes are defined subjectively by the consumer. Survey techniques can be used to assess these attribute levels that can then be entered into choice models to parameterize the intercept.

As more demand data becomes available, demand models can be used to assess the impact of marketing actions such as changes in price and advertising. Researchers are increasingly aware that standard regression and choice models are not always adequate for optimal policy determination as the parameters of these models are not policy invariant. For example, the price sensitivity parameter in a choice model can give misleading predictions if the marketing environment is altered significantly. Dynamic models of consumer choice that take into account price expectations and inventory decisions are required for policy evaluation (Gonul and Srinivasan (1996)). For example, the effects of a

temporary price cut (sale) can be very different in a market environment with frequent, predictable sales than in an environment with infrequent sales.

Dynamic considerations also affect the interpretation of the intercept parameters in choice models. In general, the consumer may update dynamically their views about product quality. For some products, past consumption and/or advertising exposure may lead to "learning" or updates of the intercept values (Erdem and Keane (1996)). In models with consumption or advertising based updating of brand preference, marketing interventions such as temporary price cuts and advertising can have long-run affects which require a dynamic structural model for policy evaluation.

In addition to dynamic considerations, we must also recognize the fact that the consumer may not be fully informed about all of the marketing mix variables when making a purchase decision. Typically, the choice models used in marketing applications assume that the consumer is aware of the prices of all products in the choice set. Studies have shown that consumers are not fully aware of prices and engage is a price search process in which they become selectively informed. This is best modeled in a search theoretic framework in which the consumer decides to sample or not sample a price of an item based on the expected return to further search. For example, if the consumer finds a relatively low price, as judged by the prior, on a preferred item, further search may have a negative expected pay-off. Estimating and postulating non-trivial choice models with price search presents an important challenge for the marketing literature (see Mehta, Rajiv and Srinivasan (1999) for a pioneering effort). A search theoretic framework can also shed light on the role of information-style advertising such as in-store displays and feature ads in local newspapers. A major role for such advertising is to inform consumers of prices without incurring the cost of search. Finally, marketers have long debated the possibility of long-term negative effects of price promotions (frequent sales). In a search model, greater price variability gives rise to a greater return on search and heightens price sensitivity. A formal search model can provide marketers with a sense of the trade-off between the long-term and short-term effects of price promotions.

Reduction in the cost of acquiring survey data will result in increased interest in statistical and psychometric methods. For example, web-based consumer satisfaction surveys present methodological challenges including the problem of respondent-specific scale usage patterns (see Yorkston (1992), and Rossi et al (1999)). More generally, survey

data provides a unique set of statistical challenges which stem from measurement error problems (see Bagozzi et al (1999) for an excellent discussion of measurement error models) as well as the multivariate aspects of the data (see DeSarbo (1994) for an overview of developments in multi-dimensional scaling). Low response rates to many marketing surveys also present a modeling and measurement challenge (see Bradlow (1999) for a recent approach to the problem of "no answer" responses).

In summary, the rapid growth in consumer data has caused a major change in the sorts of statistical models used by marketing researchers. Many standard models of demand data are often not sufficient to provide insights into consumer behavior, nor to deal with the often incomplete nature of demand data available to firms. The challenge to researchers working in marketing is to develop new statistical tools appropriate for this data environment, while employing models that can shed meaningful insight on consumer behavior.

References

Ainslie, A. and P.E.Rossi (1998) "Similarities in Choice Behavior Across Product Categories," *Marketing Science*, 17, 91-106.

Allenby, G.M.., R.P. Leone and L. Jen (1999) "A Dynamic Model of Purchase Timing with Application to Direct Marketing," *Journal of the American Statistical Association,* 94, 365-374.

Allenby, G.M. and P.E. Rossi (1999) "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, 57-78.

Allenby, G.M., N. Arora and J.L. Ginter (1998) "On The Heterogeneity of Demand," *Journal of Marketing Research*, 35, 384-389.

Allenby, G.M. and J.L. Ginter (1995) "Using Extremes to Design Products and Segment *Journal of Marketing Research*, 32, 392-403.

Allenby, G.M. and P.J. Lenk (1994) "Modeling Household Purchase Behavior with Logistic *Journal of the American Statistical Association,* 89, 1218-1231.

Bagozzi, R..P., Y. Yi and K. Nassen (1999), "Representation of Measurement Error in Marketing Variables: Review of approaches and extension to three-facet designs," *Journal of Econometrics* 89, 393-421.

Boatwright, P. and P.E.Rossi (1999) "Estimating Price Elasticities with Theory-based Priors," forthcoming, *Journal of Marketing Researach*.

Bradlow, E.T. and D.C. Schmittlein (1999) "The Little Engines That Could: Modeling the Performance of World Wide Web Search Engines," *Marketing Science*, forthcoming.

Bradlow, E.T. and A. M. Zaslavsky (1999), "A Hierarchical Latent Variable Model for Ordinal Data From a Customer Satisfaction Survey with "No Answer" Response," *JASA* 94, 43-52.

Chipman, H., E. George, and R. McCulloch (1998), "Bayesian CART Model Search," *JASA* 93, 935-960.

DeSarbo, W. S., A. K. Munrai, L. A. Munrai (1994), "Latent Class Mulit-dimensional Scaling: A review of recent developments in the marketing and psychometric literature," in Bagozzi, R. P. (ed.), *Advanced Methods for Marketing Research*, Cambridge: Blackwell.

Erdem, T. and M.P. Keane (1996) "Decision-making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets," *Marketing Science*, 15, 1-20.

Gelfand, A.E. and A.F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

Geweke, J. (1989), "Simulation Estimation Methods for Limited Dependent Variable *Handbook of Statistics, Vol 11,* G.S. Maddala, C.R. Rao and H.D. Vinod (eds), Amsterdam: North Holland.

George, E. and R. McCulloch (1995), "Variable Selection Via Gibbs Sampling," *JASA* 88, 881-889.

Gonul, F. and K. Srinivasan (1996), "Estimating the Impact of Consumer Expectations of Coupons on Purchase Behavior: A Dynamic Structural Model," *Marketing Science* 15, 262-279.

Greenleaf, E. (1992), "Improving Ratings Scale Measures by Detecting and Correcting Bias Components in Some Response Styles," *Journal of Marketing Research* 29, 176-188.

Kamakura, W. and G. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure, " *Journal of Marketing Research* 26, 379-90.

Mehta, N. S. Rajiv and K. Srinivasan (1999), "Active Versus Passive Loyalty: A Structural Model of Consideration Set Formation," working paper, Carnegie-Mellon University.

Montgomery, A.L. (1997) "Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data," *Marketing Science*, 16, 315-337.

Neelemegham, R. and P.Chintagunta (1999) "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets," *Marketing Science*, 18, 115-136.

Ripley, B.D., *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

Rossi, P.E., R.E. McCulloch and G.M. Allenby (1996) "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15, 321-340.

Rossi, P. E., Z. Gilula and G. M. Allenby (1999), "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach," working paper, Graduate School of Business, University of Chicago.