TECHNICAL WORKING PAPER SERIES

A RESEARCH ASSISTANT'S GUIDE TO RANDOM COEFFICIENTS DISCRETE CHOICE MODELS OF DEMAND

Aviv Nevo

Technical Working Paper 221
http://www.nber.org/papers/T0221

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 1998

This paper is a revised version of various chapters of my 1997 Harvard University dissertation. I wish to thank my advisors Gary Chamberlain, Zvi Griliches and Michael Whinston for guidance and support. Financial support from the Graduate School Fellowship Fund at Harvard University and the Alfred P. Sloan Doctoral Dissertation Fellowship Fund is gratefully acknowledged. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

A Research Assistant's Guide to Random Coefficients
Discrete Choice Models of Demand
Aviv Nevo
NBER Technical Working Paper No. 221
February 1998

ABSTRACT

The study of differentiated-products markets is a central part of empirical industrial organization. Questions regarding market power, mergers, innovation, and valuation of new brands are addressed using cutting-edge econometric methods and relying on economic theory. Unfortunately, difficulty of use and computational costs have limited the scope of application of recent developments in one of the main methods for estimating demand for differentiated products: random coefficients discrete choice models. As our understanding of these models of demand has increased, both the difficulty and costs have been greatly reduced. This paper carefully discusses the latest innovations in these methods with the hope of (1) increasing the understanding, and therefore the trust, among researchers who never used these methods, and (2) reducing the difficulty of use, and therefore aiding in realizing the full potential of these methods.

Aviv Nevo
Department of Economics
University of California, Berkeley
549 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
nevo@econ.berkeley.edu

1. INTRODUCTION

Much of the recent empirical industrial organization literature, the "New Empirical IO," has focused on study of a single industry (see Bresnahan, 1989.) Until the last few years, most of the literature has concentrated on homogenous-goods markets. In the last few years this trend has changed and more attention is being paid to differentiated-products markets. Questions regarding market power, mergers, innovation, and valuation of new brands are addressed using economic theory and econometric estimates of demand.[1] The econometric estimation introduces many challenges. Recent developments in the one of the main methods of dealing with these challenges, random coefficients discrete choice models, have not yet realized their full potential due, mainly, to difficulty of use. This paper carefully discusses these innovations with the intension of reducing the barriers to entry and increasing the trust in these methods among researchers that are not familiar with them.

Since the pioneering work of Stone (1954) practitioners estimating demand systems have struggled with the need for flexible functional forms, which do not impose a prior the data cannot overcome, while keeping a connection to neo-classical demand theory (either by imposing it, or finding ways to test it.) Examples include the Rotterdam model (Theil, 1965; and Barten 1966), the Translog model (Christensen, Jorgenson, and Lau, 1975), and the Almost Ideal Demand System (Deaton and Muellbauer, 1980.)

Estimating demand for differentiated products adds two additional non-trivial concerns.

---

[1]Just to mention some of these studies, Bresnahan (1987) studies the 1955 price war in the automobile industry; Gasmi, Laffont and Vuong (1992) empirically study collusive behavior in a soft-drink market; Hausman, Leonard and Zona (1994) study the beer industry; Berry, Levinsohn and Pakes (1995) examine equilibrium in the automobile industry; Goldberg (1995) uses estimates of the demand for automobiles to investigate trade policy issues; Berry, Carnall and Spiller (1996) study hubs in the airline industry; Bresnahan, Stern and Trajtenberg (1997) study rents from innovation in the computer industry; Nevo (1997b-d) examines price competition, mergers and valuation of new brands in the U.S. ready-to-eat cereal industry.

First, the large number of products, and hence the large number of parameters to be estimated. To be more specific, we require an estimate of the demand system and the pattern of substitution between the goods. If, for example, we have 200 differentiated products, then assuming constant elasticity demand curves, implies estimating 40,000 price elasticities. Even if we impose restrictions implied by economic theory, the number of parameters will still be too high to estimate with any available data set.

An additional problem, introduced when estimating demand for differentiated products, is the heterogeneity in consumer tastes: If all consumers are identical then we would not observe the level of differentiation we see in the modern marketplace. One could assume that preferences are of the "right" form (the Gorman form, see Gorman 1959), so that an aggregate, or average, consumer exists and has a demand function that satisfies the conditions specified by economic theory (for example see Dixit and Stiglitz, 1977; or Spence 1976.) However, for some applications using only the aggregate behavior has different implications from using the individual demand patterns.

A solution to both these problems is given by the discrete choice literature (for example see McFadden, 1973, 1978, 1984; Cardell, 1989; Berry, 1994; Berry, Levinsohn, and Pakes, henceforth BLP, 1995; or Nevo 1997a.) The dimensionality problem is solved by projecting the products onto a space of characteristics, making the relevant dimension the dimension of this space and not the square of the number of products. Heterogeneity is modeled explicitly and unknown parameters governing the distribution of heterogeneity are estimated. A well known problem with many discrete choice models is the strong implication of some of the assumptions made.

Recent developments in methods for estimating random coefficients discrete choice

models of demand (Cardell, 1989; Berry, 1994; BLP; and Nevo, 1997a) maintain the advantages of the discrete choice literature while (1) relaxing the restrictions on substitution patterns, (2) dealing with the endogenity of the price variable and (3) estimating the model using market-level price and quantity data. The computational costs of implementing these methods are higher than for classical discrete choice models. However, with advances in the computing methods, computation time has been reduced to reasonable levels. The rest of this article documents these methods.

A parallel line of research estimates the same models using individual data, rather than only market-level data (see Goldberg, 1995;Rossi, McCulloch and Allenby, 1996; or McFadden and Train, 1997.) This body of work differs from the one described here in the structure of the data (individual versus market) and in the assumption, usually made, that prices are exogenous (while the work described here deals with the endogenity of prices.) It should be noted that if one is using individual data the endogenity issues will not exist in many cases.

An alternative to the discrete choice methods, discussed here, is a multi-level demand model (see Hausman, Leonard, and Zona, 1994; or Hausman, 1996.) The essential idea is to use aggregation and separability assumptions to justify different levels of demand. The top level is the overall demand for the product category (for example RTE cereal.) Intermediate levels of the demand system, model substitution between various market segments, for example, between kids cereals and natural cereals. The bottom level is the choice of a brand within a segment. Each level of the demand system can be estimated using a flexible functional form. This segmentation of the market reduces the number of parameters proportionally to the inverse of the number of segments. Therefore, with either a small number of brands or a large number of (a priori) reasonable segments this method can use flexible functional forms (for example the Almost Ideal

Demand System of Deaton and Muellbauer, 1980) to give good first order approximations to any demand system. However, as the number of brands in each segment increases, beyond a handful, this method becomes less feasible. For a comparison between the methods described below and these multi-level models see Nevo (1997a, chapter 6).

The paper is organized as follows. Section 2 describes a model that encompasses, with slight alterations, the models previously used in the literature. In Section 3 I give the "big picture" of the estimation procedure. Section 4 describes the nitty-gritty details of estimation and should be read only by those readers who would like to actually code the algorithm . Section 5 concludes.

2. THE MODEL

The primitives of the model are consumer preferences and the characteristics of products. Producers and consumers in the market are assumed to observe all product characteristics and decisions of all other participants. The researcher, on the other hand, is assumed to observe only some of the product characteristics and the market outcome of price and quantity of each product. Some information regarding the distribution of consumer characteristics might be available but a key assumption is that individual consumer decisions are not observed.[2]

Formally, I assume we observe $t=1,...,T$ markets, each with $i=1,...,I_t$ consumers. For each such market we observe aggregate quantities, average prices, and product characteristics for J

[2]If such decisions are observed the method of analysis differs somewhat from the one presented here. See Goldberg (1995), Rossi, McCulloch and Allenby (1996), or McFadden and Train (1997), for ways to deal with the case where individual decisions are observed, but no aggregate data is observed. If both aggregate data and individual decisions are observed they can be combined to enrich the analysis. This is the topic of ongoing research.

products.[3] The definition of a market will vary. BLP use annual sales over a period of twenty years to define twenty different national markets. On the other hand, Nevo (1997b) defines a market as a city-quarter combination and Das, Olley and Pakes (1994) define different markets by looking at different income levels at different years.

The conditional indirect utility of consumer i from product j in market t, $U(x_{jt}, \xi_{jt}, p_{jt}, \tau_i; \theta)$, is a function of observed and unobserved (by the researcher) product characteristics, x_{jt} and ξ_{jt} respectively, price, p_{jt}, individual characteristics, τ_i and unknown parameters, θ. An important point for the econometric analysis is that the unobserved product characteristics, ξ_{jt}, are not only observed by the producers but also taken into account when setting the price. This introduces the econometric problem of endogenous prices which is similar to the well known problem of simultaneity in classical demand estimation methods. In the model presented here both prices and the unobserved characteristics influence utility in a non-linear fashion which can potentially create difficulties in a straightforward application of instrumental variables methods.

The estimation method presented below shows how to transform the model in such a way that well-known instrumental variable methods can be used. Furthermore, I explain how parameters that govern heterogenous individual demand can be estimated despite the fact that we do not observe individual behavior.

I focus on a particular specification of demand which includes, with small changes, most

---

[3]For ease of exposition I have assumed that all products are offered to all consumers in all markets. The methods described below can easily deal with the case where the choice set differs between markets and also, with somewhat more difficulty, with different choice sets to different consumers. Finally, the methods described below can also deal with cases where in some markets the quantities of several products are aggregated. For example, in some markets we might only observe the total amount of Cheerios sold (in different size boxes), yet for the purpose of our analysis we would like to define the different size boxes as different products. Of course this aggregation cannot be taken to the extreme: we cannot identify much if all we observe is the aggregation across all products. A topic for further research is identification and the extent of aggregation across products.

of the specifications used in previous work. In this specification, the conditional indirect utility of consumer i from product j in market t is[4]

$$u_{ijt} = x_{jt}\beta_i^* - \alpha_i^* p_{jt} + \xi_j + \Delta\xi_{jt} + \varepsilon_{ijt}, \qquad (1)$$
$$i=1,...,I_t \quad j=1,...,J, \quad t=1,...,T$$

where x_{jt} is a K-dimensional vector of observable characteristics of product j, p_{jt} is the price of product j in market t, ξ_j is the national mean of the unobserved (by the econometrician) product characteristics, $\Delta\xi_{jt}$ is a market specific deviation from this mean, and ε_{ijt} is a mean zero stochastic term. Finally, $(\alpha_i^* \ \beta_i^*)$ are $K+1$ individual specific coefficients.

Examples of observed characteristics vary by the product being considered. BLP examine the demand for cars, and include as observed characteristics horsepower, size and the existence of air-conditioning. In estimating demand for ready-to-eat cereal Nevo (1997a) observes calories, sodium and fiber content. Unobserved characteristics include a vertical component (for example, at equal prices all consumers weakly prefer a national brand to a generic version), components that are consumer specific (for example taste and valuation of freshness), and the market specific effects of merchandising (other than national advertising.) The reason for separating in notation ξ_j and $\Delta\xi_{jt}$ will become clear below once I discuss including brand-specific dummy variables as product characteristics.

Implicit in the specification given by equation (1) are two things. First, the indirect utility can be derived from a quasi-linear utility function, which is free of wealth effects. For some products this is a reasonable assumption, for example ready-to-eat cereals, but for other products this is a bad assumption, for example cars. Including wealth effects alters the way price and

[4]The methods discussed here are general and with minor adjustments can deal with different functional forms.

income enter equation (1). For an example of how to deal with such effects see BLP. In addition equation (1) specifies the vertical component to be identical to all consumers. Since the coefficient on price is allowed to vary among individuals this is consistent with the theoretical literature of vertical product differentiation. An alternative is to model the distribution of the valuation of the unobserved characteristics, as in Das, Olley and Pakes (1994). Note that their model is by no means more general, it only differs in the assumptions made on the distribution of the random coefficients.

The next component of the model describes how consumer preferences vary as a function of the individual characteristics, τ_i, some of which are observed. In the context of equation (1) this amounts to modeling the distribution of consumer taste parameters. The individual characteristics consist of demographics that are observed and additional characteristics that are unobserved, denoted D_i and v_i respectively. I model the distribution of consumers taste parameters for the characteristics as multi-variate normal (conditional on demographics) with a mean that is a function of demographic variables and parameters to be estimated, and a variance-covariance matrix to be estimated. Let $\gamma_i^* = (\alpha_i^*, \beta_{i1}^*, ..., \beta_{iK}^*)$ and $\gamma = (\alpha, \beta_1, ..., \beta_K)$ where K is the dimension of the observed characteristics vector; therefore,

$$\gamma_i^* = \gamma + \Pi D_i + \Sigma v_i \ , \quad v_i \sim N(0, I_{K+1}), \tag{2}$$

where D_i is a $d \times 1$ vector of demographic variables, Π is a $(K+1) \times d$ matrix of coefficients that measure how the taste characteristics vary with demographics, and Σ is a scaling matrix.[5]

The specification of equation (2) implicitly makes assumptions about both functional form and distributions. The latter is somewhat relaxed if the empirical distribution of observed

[5]Alternatively, one could think of a composite "error" term, v_i^*, which is distributed $N(0, \Sigma^*)$ and Σ is the Cholesky factorization of Σ^*.

demographics is used, as explained below. The functional form assumptions can be relaxed by including quadratic and cubic terms in D_i (for an example see Nevo, 1997b.)

The specification of the demand system is completed with the introduction of an "outside good"; the consumers may decide not to purchase any of the brands. Without this allowance a homogenous price increase (relative to other sectors) of all the products does not change quantities purchased. The indirect utility from this outside option is

$$u_{i0t} = \xi_0 + \pi_0 D_i + \sigma_0 \nu_{i0} + \varepsilon_{i0t} \ .$$

The mean utility of the outside good is not identified (without either making more assumptions or normalizing one of the "inside" goods); thus, I normalize ξ_0 to zero. The coefficients π_0 and σ_0 are not identified separately from coefficients on a constant that are allowed to vary by household. I interpret the coefficients on this constant as utility parameters of the outside good.

Let $\theta = (\theta_1, \theta_2)$ be a vector containing all the parameters of the model. The vector $\theta_1=(\alpha, \beta)$ contains the linear parameters and the vector $\theta_2=(\Pi, \Sigma, \pi_0, \sigma_0)$ the non-linear parameters[6]. Combining equations (1) and (2)

$$\begin{aligned} u_{ijt} &= \delta_{jt}(x_{jt}, p_{jt}, \xi_j, \Delta\xi_{jt}; \theta_1) + \mu_{ijt}(x_{jt}, p_{jt}, \nu_i, D_i; \theta_2) + \varepsilon_{ijt} \\ \delta_{jt} &= x_{jt}\beta - \alpha p_{jt} + \xi_{jt} + \Delta\xi_{jt}, \quad \mu_{ijt} = [p_{jt}, x_j] * (\Pi D_i + \Sigma \nu_i) \end{aligned} \tag{3}$$

where $[p_{jt}, x_j]$ is a $(K+1)\times 1$ vector. Equation (3) illustrates the key idea driving the estimation procedure described below. The utility is now expressed as the mean utility, represented by δ_{jt}, which is common to all consumers, and a mean zero heteroskedastic deviation from that mean, $\mu_{ijt} + \varepsilon_{ijt}$, that captures the effects of the random coefficients. The estimation exploits this separation to (1) reduce the number of parameters that enter in a non-linear fashion and (2) generate linear moment conditions.

---

[6]The reasons for the names will become apparent below.

Consumers are assumed to purchase one unit of the good that gives the highest utility. This implicitly defines the set of unobserved variables that lead to the choice of good j. Formally, let this set be

$$A_{jt}(x_{.t},p_{.t},\delta_{.t};\theta_2)=\left\{(D_i,v_i,\varepsilon_{it})\,|\,u_{ijt}\geq u_{ilt}\ \ \forall l=0,1,...,J\right\}$$

where x are the characteristics of all brands, $x_{.t}=(x_{1t},....,x_{Jt})'$, $p_{.t}=(p_{1t},....,p_{Jt})'$ and $\delta_{.t}=(\delta_{1t},....,\delta_{Jt})'$. Assuming ties occur with zero probability, the market share of the jth product, as a function of the mean utility levels of all the J+1 goods, given the parameters, is

$$\begin{aligned} s_{jt}(x_{.t},p_{.t},\delta_{.t};\theta_2) &= \int_{A_{jt}} dP^*(D,v,\varepsilon) = \int_{A_{jt}} dP^*(\varepsilon|D,v)\,dP^*(v|D)\,dP^*(D) \\ &= \int_{A_{jt}} dP^*(\varepsilon)\,dP^*(v)\,dP^*(D)\ , \end{aligned} \qquad \textbf{(4)}$$

where $P^*(\cdot)$ denotes population distribution functions. The second equality is a direct application of Bayes rule, while the last is a consequence of the modeling assumptions made in equations (1) and (2).

Given assumptions on the distribution of the unobserved variables we can compute the integral in equation (4), either analytically or numerically. A straightforward estimation strategy is to choose parameters that minimize the distance (in some metric) between the market shares predicted by equation (4) and the observed shares. For reasons given below, this is not the strategy used; nevertheless, it serves as an intuitive guide to the estimation scheme.

Possibly the simplest distributional assumptions one can make in order to solve the integral in equation (4) are those made in classical discrete choice models: consumer heterogeneity enters the model only through the separable additive random shock, ε_{ijt}. In our model this implies $\theta_2=0$, and therefore $\beta_{ij}^*=\beta_j$, $\alpha_i^*=\alpha$ for all i, and equation (1) becomes

$$u_{ijt}=x_{jt}\beta-\alpha p_{jt}+\xi_j+\Delta\xi_{jt}+\varepsilon_{ijt},\qquad i=1,...,I_t\quad j=1,...,J,\quad t=1,...,T. \qquad \textbf{(5)}$$

If ε_{ijt} is distributed i.i.d. with a Type I extreme value distribution, this is the well-known (Multi-nominal) Logit model. The brand market shares relative to the total market, including the outside good, are

$$s_{jt} = \frac{\exp(x_{jt}\beta - \alpha p_{jt} + \xi_j + \Delta\xi_{jt})}{1 + \sum_{k=1}^{J} \exp(x_k\beta - \alpha p_{kt} + \xi_k + \Delta\xi_{kt})} . \qquad (6)$$

Although the model implied by equation (5) and the extreme value distribution assumption is appealing, due to its tractability, it restricts the substitution patterns to depend only on the market shares. The price elasticities of the market shares defined by equation (6) are

$$\eta_{jkt} = \frac{\partial s_{jt}}{\partial p_{kt}} \frac{p_{kt}}{s_{jt}} = \begin{cases} \alpha p_{jt}(1 - s_{jt}), & \text{if } j = k; \\ -\alpha p_{kt} s_{kt}, & \text{otherwise.} \end{cases}$$

There are two problems with these elasticities. First, since in most cases the market shares are small the term $\alpha(1 - s_{jt})$ is nearly constant; hence, the own-price elasticities are proportional to own price. Therefore, the lower the price the lower the elasticity (in absolute value), which implies that a standard pricing model predicts a higher markup for the lower-priced brands. This is possible only if the marginal cost of a cheaper brand is lower (not just in absolute value, but as a percentage of price) than that of a more expensive product. For some products this will not be true. Note, that this problem is a direct implication of the functional form in price. If, for example, price enters in log form the implied elasticity would be roughly constant. In other words, the functional form directly determines the patterns of own-price elasticity.

An additional problem, which has been stressed in the literature, is with the cross-price elasticities. In the context of RTE cereals the cross price elasticities imply that if, for example, Quaker CapN Crunch (a kids cereal) and Post Grape Nuts (a wholesome simple nutrition cereal) have similar market shares, then the substitution from General Mills Lucky Charms (a kids

cereal) toward either of them will be the same. Intuitively, if the price of one kids cereal goes up we would expect more consumers to substitute to another kids cereal than to a nutrition cereal. Yet, the Logit model restricts consumers to substitute towards other brands in proportion to market shares, regardless of characteristics.

The Nested Logit (McFadden, 1978) is a slightly more complex model in which the i.i.d. extreme value assumption is replaced with a variance components structure. All brands are grouped into exhaustive and mutually exclusive sets. A consumer has a common shock to all the products in a set, so she is more likely to substitute to other products in the group. Cardell (1991) shows the distributional assumptions required in order to retain the extreme value distribution on the composite term. From a computation point of view the advantage of this model is that it implies a closed form expression for the integral in equation (4).

The Nested Logit model allows for somewhat more flexible substitution patterns, yet retains the computational simplicity of the Logit structure. In many cases the a priori division of products into groups, and the assumption of i.i.d. shocks within a group, will not be reasonable either because the division of segments is not clear or because the segmentation does not fully account for the substitution patterns. Furthermore, the Nested Logit model does not help with the problem of own-price elasticities. This is usually handled by assuming some "nice" functional form, yet does not solve the problem of having the elasticities be driven by the functional form assumption.

In some industries the segmentation of the market will be multi-layered. For example, computers can be divided into branded versus generic and into frontier versus non-frontier technology . It turns out that the results of the Nested Logit the order of the nests matters (even if the classification of the products does not change.) For this reason Bresnahan, Stern and

Trajtenberg (1997) build on the theorem by McFadden (1978) to construct a GEV model of demand for computers. With the exception of dealing with the problem of ordering the nests , this model retains all the advantages and disadvantages of the Nested Logit. In particular it implies a closed form expression for the integral in equation (4).

If in the full model, described by equations (1) and (2), we maintain the i.i.d. extreme value distribution assumption. The price elasticities of the market shares defined by equation (4) are

$$\eta_{jkt} = \frac{\partial s_{jt}}{\partial p_{kt}} \frac{p_{kt}}{s_{jt}} = \begin{cases} \dfrac{p_{jt}}{s_{jt}} \int \alpha_i s_{ijt}(1 - s_{ijt}) dP^*(D) dP^*(v), & \text{if } j = k; \\ -\dfrac{p_{kt}}{s_{jt}} \int \alpha_i s_{ijt} s_{ikt} dP^*(D) dP^*(v), & \text{otherwise.} \end{cases}$$

where $s_{ijt} = \exp\{\delta_{jt} + \mu_{ijt}\} / \left(1 + \sum_{k=1}^{K} \exp\{\delta_{jt} + \mu_{ikt}\}\right)$, is the probability of individual i purchasing the product j. Now own-price elasticity will not necessarily be driven by functional form. The partial derivative of the market shares will no longer be determined by a single parameter, α. Instead, each individual will have a different price sensitivity, which will be averaged to a mean price sensitivity using the individual specific probabilities of purchase as weights. The price sensitivity will be different for different brands. So if, for example, consumers of Kellogg's Corn Flakes have high price sensitivity, then the own-price elasticity of Kellogg' Corn Flakes will be high despite the low prices and the fact that prices enter linearly. Therefore, substitution patterns are not driven solely by functional form.

The full model also allows for flexible substitution patterns, which are not constrained by a priori segmentation of the market (yet at the same time can take advantage of this segmentation by including a segment dummy variable as a product characteristic.) The composite random shock, $\mu_{ijt} + \varepsilon_{ijt}$, is no longer independent of the product and individual characteristics. Thus, if

the price of a brand goes up consumers are more likely to switch to brands with similar characteristics, rather than to the most popular brand. Also, households with similar characteristics will tend to have similar purchasing patterns.

Unfortunately, these advantages do not come without cost. Estimation of the model specified in equation (3) is not as simple as that of the Logit, Nested Logit or GEV models. There are two immediate problems. First, equation (4) no longer has an analytic closed form (like that given in equation (6) for the Logit case.) Furthermore, the computation of the integral in equation (4) is difficult. This is solved using the simulation technique, introduced by Pakes (1986), to compute the integral. Second, we now require information on the distribution of demographics in the population in order to compute the market shares. This is solved by sampling individuals from the CPS.

3. Estimation: The Big Picture

3.1 The Data

The market-level data required to consistently estimate the model previously described consists of the following variables: market shares and prices in each market, brand characteristics, advertising, and information on the distribution of demographics. In principal, some of the parameters of the model are identified even with data on one market. However, it is highly recommended to gather data on several markets with variation in relative prices of the products and/or the products offered.

Market shares are defined using a quantity variable, which depends on the context and should be determined by the specifics of the problem. BLP use the number of automobiles sold, while Nevo (1997b) converts pounds of cereal into servings. An additional consideration in

choosing the quantity variable is the need to define a market share for the outside good. This share will very rarely be observed and will usually be defined as the total size of the market minus the shares of the inside goods. The total size of the market is assumed according to the context. So, for example, Nevo (1997b) assumes the size of the market to be one serving of cereal per capita per day. Bresnahan, Stern and Trajtenberg (1997) take the potential market to be the total number of office-based employees.

In general to following rules should be used when defining the market size. It is usually better to define the market larger rather than smaller. When looking at historical data one can use eventual growth to learn about the potential market size. Check the sensitivity of the results to the market definition; if the results are sensitive consider an alternative. There are two parts to defining the market size: choosing the variable to which the market size is proportional and choosing the proportionality factor. For example, I assumed elsewhere (Nevo 1997a) that the market size is proportional to the size of the population with the proportionality factor equal to one serving per day. It turns out that the results are not sensitive to the latter assumption, or in other words, I could define the market size as several servings per capita per day with little effect on the results[7]. The results were more sensitive to the definition of the population, but it is this part of the assumption that I felt confident about.

An important part of any data set required to implement any of the models described in Section 2 are product characteristics. These can include physical product characteristics and market segmentation information. They can be collected from manufacturer's description of the product, trade press or the researcher's prior.

In collecting product characteristics we recall the two roles in the they play in the

[7]Furthermore, the proportionality factor can be estimated (see Berry, Carnall and Spiller 1996.)

analysis: explaining the mean utility level, $\delta(\cdot)$ in equation (3), and driving the substitution patterns through the term $\mu(\cdot)$ in equation (3). Ideally, these two roles should be kept separate. If the number of markets is large enough, relative to the number of products, the mean utility can be explained by including product dummy variables in the regression. These variables will absorb any product characteristics that are constant across markets. A discussion of the issues arising from including brand dummy variables is given below.

Relying on product dummy variables to guide substitution patterns is equivalent to estimating an unrestricted variance-covariance matrix of the random shock, ε_{ijt} in equation (1). Both imply estimating J^2 parameters. Since part of our original motivation was to reduce the number of parameters to be estimated, this is usually not a feasible option (see Hausman and Wise 1978, for such a model with a small number of products) . The substitution patterns are explained by the product characteristics and in deciding which attributes to collect the researcher should keep this in mind.

The last component of the data is information regarding the demographics of the consumers in different markets. Unlike market shares, prices or product characteristics, the estimation can proceed without demographic information. Estimation, in such a case, will rely on assumed distributional assumptions rather than empirical distributions. The Current Population Survey (CPS) is a good widely available source for demographic information.

3.2 Identification

This section discuses informally and intuitively some of the identification issues. A useful starting point is to ask how one would approach the problem of estimating price elasticities if a controlled experiment could be conducted. The answer is to expose different

consumers to randomly assigned prices and record their purchasing patterns. Furthermore, one could relate these purchasing patterns to individual characteristics. If individual purchases could not be observed, the experiment could still be run by comparing the different aggregate purchases of different groups. Once again these patterns could be related to the difference in individual characteristics between the groups.

There are two potential problems with mapping the data previously described into the data that arises from the ideal controlled experiment. First, prices are not randomly assigned, rather, they are set by profit maximizing firms that take into account information that due to inferior knowledge the researcher has to include in the error term. This problem will be solved below by using instrumental variables.

The second, somewhat more conceptual difficulty, arises because discrete choice models, for example the Logit model, can be estimated using data from just one market; hence, we are not mimicking the experiment previously described. Instead in our experiment we ask consumers to choose between products, which are perceived as bundles of attributes. We then reveal the preferences for these attributes which include price. The data from each market should not be seen as one observation of purchases when faced with a particular price vector, rather, it is an observation on the relative likelihood of purchasing J different bundles of attributes. The discrete choice model ties these probabilities to a utility model which allows us to identify price elasticities. The identifying power of this experiment increases as more markets are included with variation both in the characteristics of products and the choice set. The same (informal) identification argument holds for the Nested Logit, GEV, and random coefficients models, which

are just a generalized form of the Logit model.[8]

There are two caveats to the informal argument previously given. If one wants to tie demographic variables to observed purchases (i.e., allow for a ΠD_i in equation (2)) several markets, with variation in the distribution of demographics, have to be observed. Second, if not all the product characteristics are observed and these unobserved attributes are correlated with some of the observed characteristics, then we are faced with an endogenity problem. The problem can be solved by using instrumental variables, but we note that the formal requirements from these IV's is different than those required earlier in this section. I return to this point in Section 3.4. An alternative to IV methods is the inclusion of brand-specific dummy variables that I strongly recommend if enough markets are observed.

In summary, this section outlined an informal identification argument. The key to consistent estimation lies in finding valid instrumental variables (in a sense that will be made clear in the next section.) The requirements from these IV's are different if several markets are observed and brand-specific dummy variables are included as brand characteristics. Furthermore, the experiment that we are trying to reproduce is different.

3.3 The Estimation Algorithm

In this section I outline how the parameters of the models described in Section 2 can be consistently estimated using the data described in the Section 3.1. I follow the suggestion of Berry (1994) to construct a GMM estimator. Essentially, the idea is to compute the implied "error term", for a given value of the unknown parameters, and interact it with instruments, thus

[8]In principal, a random coefficient model, like the one used by BLP, should be identified with observations on only one market. In practice, however, BLP report difficulty estimating the model with a single market (see BLP footnote 30.) This should not be surprising given the previous comment on the importance of variation in the choice set.

forming the GMM objective function. Next, a search is performed over all the possible parameter values to find those values that minimize the objective function. In this section I discuss what the error term is, how it can be computed, and some computational details. Discussion of the instrumental variables is deferred to the next section.

As previously pointed out, a straightforward approach to the estimation is to solve

$$\underset{\theta}{\text{Min}} \left\| s(x,p,\delta;\theta) - S \right\| \tag{7}$$

where $s(\cdot)$ are the market shares given by equation (4) and S are the observed market shares. However, this approach will not be taken for several reasons. First, all the parameters enter the minimization in equation (7) in a non-linear fashion. In some applications the inclusion of brand and time dummy variables results in a large number of parameters and a costly non-linear minimization problem. The estimation procedure suggested by Berry (1994), which is used below, avoids this problem by transforming the minimization problem such that some (or all) of the parameters enter the objective function linearly. Furthermore, it is much harder to think of identifying assumptions in the context of equation (7), while it is totally natural to do so in the method used below.

Formally, let $Z=[z_1,\ldots,z_M]$ be a set of instruments such that

$$E\left[Z \cdot \omega(\theta^*)\right] = 0 \ , \tag{8}$$

where ω, a function of the model parameters, is an "error term" defined below and θ^* denotes the "true" value of these parameters. The GMM estimate is

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \ \omega(\theta)' Z A^{-1} Z' \omega(\theta) \tag{9}$$

where A is a consistent estimate of $E[Z'\omega\omega'Z]$. The logic driving this estimate is simple enough. At the true parameter value (θ^*), the population moment, defined by equation (8), is equal to zero, so we choose our estimate such that it sets the sample analog of the moments

defined in equation (8), i.e. $Z'\hat{\omega}$, to zero. If there are more independent moment equations than parameters (i.e., dim(Z)>dim(θ)), we can not set all the sample analogs exactly to zero and will have to set them as close to zero as possible. The weight matrix, A, defines the metric by which we measure how close to zero we are. By using the variance-covariance matrix of the moments, we give less weight to those moments (equations) that have a higher variance.

Following Berry (1994), the "error term" is not defined as the difference between the observed and predicted market shares; rather it is obtained by inverting the market share function to obtain the vector of mean valuations that equates the observed market shares to the predicted shares. This is done by solving, for each market, the implicit system of equations

$$s_{.t}(\delta_{.t};\theta_2) = S_{.t} \ .$$

In some cases (for example, the Logit or Nested Logit models) this can be solved analytically. However, for the full model suggested above, this has to be done numerically. Once this inversion has been done, either analytically or numerically, the "error term" is defined as

$$\omega_{jt} = \delta_{jt}(S_{.t};\theta_2) - (x_j\beta + \alpha p_{jt}) \ . \qquad \textbf{(10)}$$

Note, that it is the observed market shares, S, that enter this equation. Also, we can now see the reason for distinguishing between θ_1 and θ_2: θ_1 enters this term, and the GMM objective, in a linear fashion, while θ_2 enters non-linearly.

The intuition in this definition is as follows. For a given value of the non-linear parameters, θ_2, we compute what is the mean valuation, $\delta_{jt}(\cdot)$, that would set the predicted market share equal to the observed market share. We define the residual as the difference between this valuation and the one "predicted" by the linear parameters, α and β. The estimator, defined by equation (9), is the one the minimizes the distance between these different predictions.

Usually[9], the error term, as defined by equation (10), is the unobserved product characteristic, ξ_j. However, if enough markets are observed then brand-specific dummy variables can be included as product characteristics. The coefficients on these dummy variable include both the mean quality index of observed characteristics, βx_j, and the unobserved characteristics, ξ_j. Thus, the error term is the market specific deviation from the main valuation, i.e., $\Delta\xi_{jt}$. The inclusion of brand dummy variables introduces a challenge in estimating the taste parameters, β, which is dealt with below.

In the Logit and Nested Logit models, with the appropriate choice of a weight matrix[10], this procedure simplifies to two-stage least squares. In the full random coefficients model, both the computation of the market shares, and the "inversion" in order to get $\delta_{jt}(\cdot)$, have to be done numerically. The value of the estimate in equation (9) is then computed using a non-linear search. This search is simplified by noting that the first order conditions of the minimization problem defined in equation (9) with respect to θ_1 are linear in these parameters. Therefore, these linear parameters can be solved for (as a function of the other parameters) and plugged into the rest of the first order conditions, limiting the non-linear search to only the non-linear parameters.

The details of the computation are given in Section 4.

3.4 Instruments

The key identifying assumption in the algorithm previously given is equation (8), which requires a set of exogenous instrumental variables. The first set that comes to mind are the

---

[9]See for example Berry (1994), BLP (1995), Berry, Carnall, and Spiller (1994), Bresnahan, Stern and Trajtenberg (1997).

[10]I.e., $A=Z'Z$, which is the "optimal" weight matrix under the assumption of homoscadestic errors.

instrumental variables defined by ordinary least squares, namely the regressors (or more generally the derivative of the moment function with respect to the parameters.) In order to determine the validity of this assumption we examine the pricing decision that results from a wide variety of pricing models. Prices are a function of marginal costs and a markup term,

$$p_{jt} = mc_{jt} + f(\xi_{jt},...) = (mc_j + f_j) + (\Delta mc_{jt} + \Delta f_{jt}) \quad . \tag{11}$$

The markup term is a function of the unobserved product characteristics, which is also the error term in the demand equation (if product-specific dummy variables are not included.) Therefore, prices will be correlated with the error term and the estimate of the price sensitivity, α, biased towards zero.

Alternatively, if brand dummy variables are included in the regression the error term is the unobserved market specific deviation from the mean valuation of the brand. Since I assumed that players in the industry observe and account for this deviation, it will influence the market specific markup term and once again downward bias the estimate of price sensitivity. Indeed, the results given in BLP and Nevo (1997a) support this conclusion.

When brand dummy variables are not included in the product characteristics, the instrumental variables used in much of the literature are the product characteristics (excluding price and other potentially endogenous variables), the sum over all the firm's other brand of the same characteristics (if the firm produces more than one product) and the sum over all the firm's competing brands characteristics. The assumption justifying these IV's is that the unobserved characteristics are mean independent of the observed characteristics (see BLP.) Intuitively, they attempt to proxy for the degree and closeness of competition the brand is facing.

Once a brand dummy is introduced a problem arises with these IV's: unless there is variation in the products offered in different markets, there is no variation between markets in

these instruments. For this reasons Nevo (1997b) uses two sets of instruments in an attempt to separate the exogenous variation in prices (due to differences in marginal costs) and endogenous variation (due to differences in unobserved valuation.) First, the panel structure of the data is exploited. The identifying assumption is that, controlling for brand specific means and demographics, city specific valuations are independent across cities (but are allowed to be correlated within a city over time[11].) Given this assumption, the prices of the brand in other cities are valid instruments; from equation (11) we see that prices of brand j in two cities will be correlated due to the common marginal cost, but due to the independence assumption will be uncorrelated with market specific valuation. Potentially, one could use the prices in all other cities and all quarters as instruments.

There are several plausible situations in which the independence assumption will not hold. Suppose there is a national (or regional) demand shock, for example discovery that fiber reduces the risk of cancer. This discovery will increase the unobserved valuation of all fiber intensive cereal brands in all cities, and the independence assumption will be violated.

Suppose one believes that local advertising and promotions are coordinated across city borders, but within regions, and that these activities influence demand. Then the independence assumption will be violated for cities in the same region, and prices in cities in the same region will not be valid instruments.

Determining how plausible are these, and possibly other, situations is an empirical issue so as an alternative Nevo (1997b) examines another set of instruments that attempts to proxy for the marginal costs directly and compares the difference between the estimates implied by the

[11]This assumption is also made by Hausman (1996), although his setup does not fit into the model discussed here.

different sets of instruments.

BLP discuss the use of "optimal" IV which arises from their assumption of conditional mean independence, i.e., $E(\omega | Z) = 0$, rather than no correlation assumption, $E(Z\omega) = 0$, made here. Their assumption, which is stronger than the assumption made here, implies that not only are Z valid IV's, but so are many other functions of Z. The question they try to address is which of these functions should be used. The bottom line of their discussion is an efficiency argument as justification for the use of the first set of instruments discussed in this section.

3.5 Brand Specific Dummy Variables

As previously pointed out, I believe that brand specific fixed effects should be used whenever possible. There are at least two good reasons to include these dummy variables. First, in any case where we are unsure that the observed characteristics capture the true factors that determine utility, fixed effects should be included in order to improve the fit of the model. We note this helps fit the mean utility level, δ_j, while substitution patterns are driven by observed characteristics (either physical characteristics or market segmentation), as is the case if we were not to include a brand fixed effect.

Furthermore, a major motivation (see Berry, 1994) for the estimation scheme previously described is the need to instrument for the correlation between prices and the unobserved quality of the product, ξ_j. A brand specific dummy variable captures the characteristics that do not vary by market, namely, $x_j\beta + \xi_j$. Therefore, the correlation between prices and the unobserved quality is fully accounted for and does not require an instrument. In order to introduce a brand dummy variable we require observations on more than one market. However, even without brand dummy variables, fitting the model using observations from a single market is difficult (see BLP,

footnote 30.)

Once brand dummy variables are introduced, the error term is no longer the unobserved quality. Rather, it is market specific deviation from this unobserved mean. This additional variance was not introduced by the dummy variables; it is present in all models that use observations from more than one market. The use of brand dummy variables forces the researcher to discuss this additional variance explicitly.

There are two potential objections to the use of brand dummy variables. First, as previously mentioned, a major difficultly in estimating demand in differentiated product markets is that the number of parameters increases proportionally to the square of the number of products. The main motivation for the use of discrete choice models was to reduce this dimensionality problem. Does introduction of parameters that increase in proportion to the number of brands defeat the whole purpose?

No. The number of parameters increases only with J (the number of brands) and not J^2. Furthermore, the brand dummy variables are linear parameters, and do not increase the computational difficulty. If the number of brands is large the size of the design matrix might be problematic, but given the computing power required to run the full model this problem seems meager.

A more serious objection to the use of brand dummy variables is that taste coefficients, β, cannot be identified . Fortunately, this is not true. The taste parameters can be retrieved by using a minimum distance procedure (as in Chamberlain, 1982.)

Let $d=(d_1, ..., d_J)'$ denote the $J\times 1$ vector of brand dummy coefficients, X be the $J\times K$ ($K<J$) matrix of product characteristics which are fixed across markets, and $\xi=(\xi_1, ..., \xi_J)'$ be the $J\times 1$ vector of unobserved product qualities. Then from equation (1)

$$d = X\beta + \xi \ .$$

The estimates of β and ξ are

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\hat{d}, \quad \hat{\xi} = \hat{d} - X\hat{\beta}$$

where $\hat{d}$, is the vector of coefficients estimated from the procedure described in the previous section, and Ω is the variance-covariance matrix of these estimates. The coefficients on the brand dummy variables provide an "unrestricted" estimate of the mean utility. The minimum distance estimates project these estimates onto a lower dimensional space, which is implied by a "restricted" model that sets ξ to zero. Chamberlain provides a χ^2 test to evaluate these restrictions.

4. Estimation: The Nitty-Gritty

In this section I discuss the computational details of the estimation algorithm previously described. I describe the actual implementation of the algorithm, but also mention possible pitfalls and ways to speed the computation. The actual implementation varies for different models, differences are pointed out.

Before going into the details we recall the intuition of the estimation algorithm. Our model is one of individual behavior, yet we only observe aggregate data. Nevertheless, we can still estimate the parameters that govern the distribution of individuals by computing predicted individual behavior and aggregating over individuals, for a given value of the parameters, in order to obtain predicted market shares. We then choose the values of the parameters that minimize the distance between these predicted shares and the actual observed shares. As explained in Section 3, the metric under which this distance is minimized is not the straightforward sum of least squares, rather it is the metric defined by the instrumental variables

and the GMM objective function. It is this last step that somewhat complicates the estimation procedure.

There are essentially four steps (plus an initial step) to follow in computing the estimates:

(0) prepare the data including draws from the distribution of ν and D (only some models);

(1) for a given value of θ_2 and δ, compute the market shares implied by equation (4) (only for some models);

(2) for a given θ_2, compute the vector δ that equates the market shares computed in Step 1 to the observed shares;

(3) for a given θ, compute the error term (as a function of the mean valuation computed in Step 2), interact it with the instruments, and compute the value of the objective function;

(4) search for the value of θ that minimizes the objective function computed in Step 3.

Step 0: The data required is described in detail in Section 3.1. The actual organization of the data depends on the code used to compute the estimation. I found it useful to define a vector of market shares and two matrices of "right-hand side" variables, X_1 and X_2. The first contains the variables that enter the linear part of the estimation, common to all individuals ($\delta(\cdot)$ in equation (3)), and includes: price, advertising, brand dummy variables (or product characteristics if brand dummy variables are not included.) The latter, X_2, contains the variables that will have a random coefficient, and therefore will enter the non-linear part ($\mu(\cdot)$ in equation (3)), and includes: price, and characteristics (which could include segment-specific dummy variables.) Some models (e.g., Logit, Nested Logit, and GEV) do not have random coefficients, the non-linear part is set to zero, and therefore only the first design matrix, X_1, is required.

Next, we need to sample *ns* individuals, which for our purposes consists of a vector of

shocks that determine the individual's taste parameter, ν, demographics, D, and potentially a shock to the utility, ε. The use of this "individual data" is explained below, in Step 1. In most cases we will not need to draw ε, since we can integrate the extreme value shocks analytically (see details below.) That is also the reason we do not need these draws for the Logit, Nested Logit and GEV models.

The sample of individuals can be generated in two ways: by assuming a parametric functional form, or using the empirical non-parametric distribution of real individuals. Shocks that determine the individuals taste parameters, ν, are drawn from a multi-variate normal distribution. In principal these could be drawn from any other parametric distribution[12]. The choice of the distribution depends on the problem at hand and on the researcher's prior. Demographics are drawn from the CPS; instead of assuming parametric forms for the distribution of demographics real individuals are sampled (and if we observed actual purchases made by these individuals, we would not need the estimation steps that follow.) This, of course, has the advantage of not depending on a somewhat arbitrary assumption of a parametric distributional form. On the other hand, if the parametric form was correct (or nearly so), then using the non-parametric approach will be significantly less efficient. The choice of parametric vs. non-parametric form should depend on the researcher's believe on how well a parametric function can explain the distribution of heterogeneity.

As pointed out, the data consists of observations on prices and market shares of different brands in several markets. In theory we could consider drawing different individuals for each observation, i.e., each brand in each market. However, in order for Step 2 to work we require that the predicted market shares sum up to one, which requires that we use the same draws for

[12] For example, Berry, Carnall and Spiller (1996) use a bimodal distribution.

each market.

There is a question of whether these draws should vary between markets. Once again, the answer depends on the specifics of the problem and the data. In general, the draws should be the same whenever it is the same individuals making the decisions. For example, BLP use national market shares of different brands of cars over twenty years. They use the same draws for all markets. On the other hand Nevo (1997b) allows for different draws for each market.

Finally, it is important to draw these only once at the beginning of the computation. If the draws are changed during the computation the non-linear search is unlikely to convergence.

Step 1: Now that we have a sample of individuals (each described by (v, D, ε)), for a given value of the parameters, θ_2, and the component common to all consumers, δ, we compute the predicted market shares given by the integral in equation (4). For many models (e.g., Logit, Nested Logit and GEV) this step can be performed analytically.

For the full random coefficients model this integral has to be computed by simulation. There are several ways to do this. The first is the naive frequency estimator. We sample *ns* draws of the vector (v, D, ε), for each of these we predict the brand purchased, and then sum over the purchases. Formally, our estimator is given by

$$s_j(p,x,\delta,P_{ns};\theta_2) = \frac{1}{ns}\sum_{i=1}^{ns} 1\left\{u_{ij}(D_i,v_i,\varepsilon_i) \geq u_{il}(D_i,v_i,\varepsilon_i) \;\; \forall l=0,1,\ldots,J\right\},$$

where $1\{C\} = 1$ if the C is true, and 0 otherwise. This is repeated for every brand in every market.

Although this approach is very intuitive it is lacking in three dimensions. First, it requires a large number of draws, *ns*, at least the number of products, and usually much more than that. Otherwise, the predicted market shares for some of the products will be zero. The

smaller the observed shares the larger the problem. Second, in this approach there are three simulation processes that induce variance in the estimated parameters, the sampling processes of the v's, the D's, and the ε's. While in the approach suggested below the simulation error due to the sampling process of the ε's is eliminated. Finally, this approach renders a non smooth objective function, and therefore does not allow us to use gradient methods (see Step 4.)

A second approach to computing the predicted market shares, which does not suffer from the above problems, is the smooth simulator. Here we use the extreme value distribution on $P^*(\varepsilon)$ to integrate the ε's analytically. Formally, the predicted market shares, given by equation (4), are approximated by

$$s_j(p,x,\delta,P_{ns};\theta_2) = \frac{1}{ns}\sum_{i=1}^{ns} s_{ji} =$$

$$\frac{1}{ns}\sum_{i=1}^{ns} \frac{\exp\left(\delta_j + \sum_{k=1}^{K} x_{jk}(\sigma_k v_{ik} + \pi_{kI} D_{iI} + \ldots + \pi_{kd} D_{id})\right)}{1 + \sum_{m=1}^{J} \exp\left(\delta_m + \sum_{k=1}^{K} x_{mk}(\sigma_k v_{ik} + \pi_{kI} D_{iI} + \ldots + \pi_{kd} D_{id})\right)},$$

where $(v_{i1},\ldots,v_{iK})$ and $(D_{i1},\ldots D_{id})$, $i=1,\ldots,ns$, are the draws from step 0.

Note, that we no longer require a large number of draws to predict non-zero market shares. Actually, one draw of the pair (v, D) will suffice (although this is by no means the recommended number.) Also, since the ε's are integrated analytically the variance due to the simulation process is limited only to the simulation of v and D. Finally, unlike before the predicted market shares are smooth functions of the parameters, and therefore a gradient method can be used to minimize the objective function.

An additional approach that reduces even further the simulation variance is the use of importance sampling (see BLP; or Gweke, 1995.)

Step 2: As previously pointed out, we choose our estimates by minimizing the distance

of the predicted market shares to the observed market shares. For the reasons given in Section 3 we will not define the distance as the sum of squares, rather we obtain an expression that is linear in the structural error term and interact it with instruments to form a GMM objective function. This step is required in order to obtain this structural error term.

We want to compute the J-dimensional vector of mean valuations, δ, that equates the market shares computed in Step 1 to the observed shares. This amounts to solving the system of equations[13]

$$s(\delta;\theta_2) = S, \tag{12}$$

where $s(\cdot)$ are the predicted market shares computed in Step 1 and S are the observed market shares. This system is solved market by market.

For the Logit model this inversion can be computed analytically by $\delta_j = \ln(S_j) - \ln(S_0)$, where S_0 is the market share of the outside good, computed by subtracting the sum of observed market shares of all the inside goods from 1. Note, it is the observed market shares that enter this equation, and therefore we do not require Step 1 for the Logit model. This inversion can also be computed analytically in the Nested Logit model (see Berry, 1994) and the GEV model (see Bresnahan, Stern and Trajtenberg, 1997.)

For the full model the system of equations (12) is non-linear and is solved numerically. I use the contraction mapping suggested by BLP (see there for proof of convergence), which amounts to computing

$$\delta^{t+1} = \delta^t + \ln(S) - \ln\left(s(p,x,\delta^t,P_{ns};\theta_2)\right), \quad t = 0,\ldots,T, \tag{13}$$

where $s(\cdot)$ are the predicted market shares computed in Step 1, T is the smallest integer such that $\|\delta^T - \delta^{T-1}\|$ is smaller than some tolerance level, and δ^T is the approximation to δ.

---

[13]Berry (1994) proves existence and uniqueness of the vector δ.

A few practical points. First, convergence can be reached faster by choosing a good starting value, δ_0. I used the value, δ^T, computed from the last iteration of the objective function, with the first iteration using the values that solve the Logit model. Better approximations can be derived (for example, using the Jacobian of $\delta(\cdot)$ computed below, the first term of a Taylor approximation can be computed.) However, I found that computing better approximations usually took longer than the time saved.

Second, the further away from the true value of the parameters we are, the harder it is to solve the system of equations (12). Therefore, I found it helpful to change the tolerance level as the number of iterations in the process defined by equation (13) goes up. For example, for the first 100 iterations, of the process defined in equation (13), the tolerance level is set to 10E-8, while after that every additional 50 iterations the tolerance level increases by an order of magnitude. If the change in the value of the parameters was low (say less than 0.01) the tolerance level was kept at its initial value. This promises that near the solution the inversion is computed very accurately, while the algorithm does not waste time far away from the solution, where the number of iterations required to achieve accurate convergence of the process defined in equation (13) is high.

Third, a trivial point but one that can save several days of work: the order of the terms in equation (13) matters.

Step 3: Finally we can compute the error term. For both the Logit and the full model it is defined by

$$\omega = \delta - X_1\theta_1,$$

where δ is the J-dimensional vector computed in Step 2, X_1 is the design matrix defined in Step 0, and θ_1 is the vector of linear parameters. This is interacted in a straightforward way with the

instrument matrix to form the moment conditions, which are used to compute the objective function

$$\omega(\theta)'ZA^{-1}Z'\omega(\theta),$$

where A is a consistent estimate of $E[Z'\omega\omega'Z]$.

Like many GMM estimation problems, the computation of the objective function requires knowledge of the weight matrix, which in general requires knowledge of the true value of the parameters (or a consistent estimate thereof.) There are several solutions to this problem. First, assume homoscedastic errors and therefore the optimal weight matrix is proportional to $Z'Z$, thus, the need to know the true parameter values is eliminated. Second, we can compute an estimate of θ, say $\hat{\theta}$, using $A = Z'Z$, and then use this estimate to compute a new weight matrix $E[Z'\omega(\hat{\theta})\omega(\hat{\theta})'Z]$, which in turn is used to compute a new estimate of θ.

An additional option is to continue iterating between the estimates of θ and A until convergence. Hansen, Heaton and Yaron (1996) show this is equivalent to solving

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ \omega(\theta)'Z\left[Z'\omega(\theta)\omega(\theta)'Z\right]^{-1}Z'\omega(\theta),$$

and title it continuous updating. Although this approach has some appeal, it is not clear in what sense it is optimal. Both this approach and the previous one are first- order asymptotically equivalent. As shown by Hansen, Heaton and Yaron (1996), for linear GMM problems, neither method dominates in finite samples. Therefore, I see no justification to use continuous updating in non-linear GMM problems, like the one here, where the computational costs are enormous

Step 4: Search for the value of θ that minimizes the objective function. For the Logit model this can be done analytically (it is a linear GMM problem.) For the full model we need to perform a non-linear search over θ. This search can be sped by using the first order conditions, with respect to θ_1, to express θ_1 as a function of θ_2, i.e.,

$$\hat{\theta}_1 = (X_1' Z A^{-1} Z' X_1)^{-1} X_1' Z A^{-1} Z' \delta(\hat{\theta}_2),$$

where X_1 is the design matrix and Z is the instrument matrix. Now, the non-linear search can be limited to θ_2.

One of two search methods is usually used. The Nelder-Mead (1965) non-derivative "simplex" search method, and a quasi-Newton method with an analytic gradient (see Press et al., 1994.) The first is more robust but is much slower to converge, while the latter is two orders of magnitude faster, yet is sensitive to starting values; due to the non-linear objective function. If the initial values are extremely poor the algorithm would reach regions where the objective is not defined. This is especially true if the variables are scaled differently, i.e., one characteristic will range from 1 to 5 while another from 0 to 250.

A useful trick is to set the value of the objective function to a high number (say 10E+10) if one of its components is not defined. This allows the algorithm to deal with situations in which the objective function is not defined. Poor starting values, different scaling of the variables, and the non-linear objective would cause this to happen. This usually happens in the first few iterations, especially if the parameters are not properly scaled. An alternative way of dealing with this problem is to decrease the initial step size (i.e., the step taken in the direction determined by the negative of the gradient.) The recommended practice is to start with the non-derivative method and switch to the gradient method once the objective has been lowered to reasonable levels.

Always when performing a non-linear search one has to assure that a global minimum was found rather than a local one. The non-derivative method is slightly more robust to local minimums because it does not use local (derivative) information. However, it also does not guarantee convergence to a global minimum. I recommend the following check list. After a

minimum is found using the gradient method, restart the problem using this minimum as the initial value for a non-derivative search. Hopefully, the results will be the same. Next, restart the problem with several different initial values and use the gradient method. Hopefully, the same results are obtained regardless of the starting values.

The main increase in speed using the gradient method is due to the analytic gradient. Using a gradient computed from finite differences results in relatively less increase in speed, while maintaining all the problems of the gradient method. In order for the gradient to be well defined everywhere, the simulated market shares have to be smooth functions of the parameters. The smooth simulator discussed above has this property, while the naive frequency simulator does not.

In order to compute the gradient of the objective function, the Jacobian of the function computed in Step 2 has to be computed. The mean valuations of the J brands in each market are implicitly defined by the following system of J equations

$$s_j(\delta_1,...,\delta_J,\theta_2;x,p,P_{ns}) = S_j, \quad j = 1,...,J.$$

By the Implicit Function Theorem (see Simon and Blume, 1994, Theorem 15.7 pg 355) the derivatives of the mean value with respect to the parameters are

$$D\delta = \begin{pmatrix} \frac{\partial \delta_1}{\partial \theta_{21}} & \cdots & \frac{\partial \delta_1}{\partial \theta_{2L}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \delta_J}{\partial \theta_{21}} & \cdots & \frac{\partial \delta_J}{\partial \theta_{2L}} \end{pmatrix} = - \begin{pmatrix} \frac{\partial s_1}{\partial \delta_1} & \cdots & \frac{\partial s_1}{\partial \delta_J} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_J}{\partial \delta_1} & \cdots & \frac{\partial s_J}{\partial \delta_J} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial s_1}{\partial \theta_{21}} & \cdots & \frac{\partial s_1}{\partial \theta_{2L}} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_J}{\partial \theta_{21}} & \cdots & \frac{\partial s_J}{\partial \theta_{2L}} \end{pmatrix}, \tag{14}$$

where θ_{2i}, $i = 1,...,L$ denotes the i's element of the vector θ_2, which contains the non-linear parameters of the model. The share function defined by the smooth simulator is

$$s_j = \frac{1}{ns}\sum_{i=1}^{ns} s_{ji} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\exp\left(\delta_j + \sum_{k=1}^{K} x_{jk}(\sigma_k v_{ik} + \pi_{k1} D_{i1} + \ldots + \pi_{kd} D_{id})\right)}{1 + \sum_{m=1}^{J} \exp\left(\delta_m + \sum_{k=1}^{K} x_{mk}(\sigma_k v_{ik} + \pi_{k1} D_{i1} + \ldots + \pi_{kd} D_{id})\right)} .$$

Therefore, the derivatives are

$$\frac{\partial s_j}{\partial \delta_j} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\partial s_{ji}}{\partial \delta_j} = \frac{1}{ns}\sum_{i=1}^{ns} s_{ji}(1 - s_{ji})$$

$$\frac{\partial s_j}{\partial \delta_m} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\partial s_{ji}}{\partial \delta_m} = -\frac{1}{ns}\sum_{i=1}^{ns} s_{ji} s_{mi}$$

$$\frac{\partial s_j}{\partial \sigma_k} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\partial s_{ji}}{\partial \sigma_k} = \frac{1}{ns}\sum_{i=1}^{ns} s_{ji}\left(x_{jk} v_{ik} - \sum_{m=1}^{J} x_{mk} v_{ik} s_{mi}\right) = \frac{1}{ns}\sum_{i=1}^{ns} v_{ik} s_{ji}\left(x_{jk} - \sum_{m=1}^{J} x_{mk} s_{mi}\right)$$

$$\frac{\partial s_j}{\partial \pi_{kd}} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\partial s_{ji}}{\partial \pi_{kd}} = \frac{1}{ns}\sum_{i=1}^{ns} s_{ji}\left(x_{jk} D_{id} - \sum_{m=1}^{J} x_{mk} D_{id} s_{mi}\right) = \frac{1}{ns}\sum_{i=1}^{ns} D_{id} s_{ji}\left(x_{jk} - \sum_{m=1}^{J} x_{mk} s_{mi}\right)$$

Substituting this back into equation (14), we obtain the Jacobian of the function computed in Step 2.

The gradient of the objective function is

$$2 * D\delta' * Z * A^{-1} * Z' * \omega .$$

Once this is programed it is easy to check, by comparing the results to the gradient computed using finite differences. If performing such a comparison, then the tolerance levels in Step 2 have to be set low (10E-8), and not allowed to vary (see discussion above.)

5. Concluding Remarks

As was mentioned in Sections 3.2-3.4, the identification relies on having a set of exogenous instrumental variables. Finding such IV's is crucial for any consistent estimation of demand. In estimating demand for differentiated-products this problem is further complicated. Cost data are rarely observed and indicators closely related to cost will rarely exhibit enough

cross-brand variation. A few solutions previously used were offered, yet all suffer from potential drawbacks. It is important not to get carried away in the technical "fireworks" and remember this most basic, yet very difficult, identification problem.

This paper mentioned some of the previous work that used the methods described here. The scope of application and potential of use are far from exhausted. Of course, there are many more potential applications within the study of industrial economics, both in studying new industries and in answering different questions. However, the full scope of these methods is not limited to industrial organization. Fields like economics of crime, public finance or environmental economics, are just a few examples of fields in which there seems to be a lot of individual heterogeneity yet individual data is rare. The model and the method described here could be adopted to deal with questions in these fields.

REFERENCES

Barten, A.P. (1966), *Theorie en Empirie van een Volledig Stelsel van Vraagvergelijkingen*, Doctoral dissertation, Rotterdam: University of Rotterdam.

Berry, S. (1994), "Estimating Discrete-Choice Models of Product Differentiation," *Rand Journal of Economics*, 25, 242-262.

Berry, S., M. Carnall, and P. Spiller (1996), "Airline Hubs: Costs and Markups and the Implications of Consumer Heterogeneity," National Bureau of Economic Research, Working paper no. 5561 (also available at http://www.econ.yale.edu/~steveb/imdex.html)

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841-890.

Bresnahan, T. (1987), "Competition and Collusion in the American Automobile Oligopoly: The 1955 Price War," *Journal of Industrial Economics*, 35, 457-482.

Bresnahan, T. (1989), "Empirical Methods for Industries with Market Power," in R. Schmalensee and R. Willig, eds., *Handbook of Industrial Organization*, Vol. II, Amsterdam: North-Holland.

Bresnahan, T., S. Stern, and M. Trajtenberg (1997), "Market Segmentation and the Sources of Rents from Innovation: Personal Computers in the Late 1980's," *RAND Journal of Economics*, 28.

Cardell, N.S. (1989), *Extensions of the Multinomial Logit: The Hedonic Demand Model, The Non-Independent Logit Model, and the Ranked Logit Model*, Ph.D. Dissertation, Harvard University.

Cardell, N.S. (1991), "Variance Components Structures for the Extreme Value and Logistic Distributions," mimeo, Washington State University.

Chamberlain, G. (1982), "Multi Variate Regression Models for Panel Data," *Journal of Econometrics*, 18(1), 5-46.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau (1975), "Transcendental Logarithmic Utility Functions," *American Economic Review*, 65, 367-83.

Das, S., S. Olley, and A. Pakes (1994), "Evolution of Brand Qualities of Consumer Electronics in the U.S.," mimeo.

Deaton, A., and J. Muellbauer (1980), "An Almost Ideal Demand System," *American Economic Review,* 70, 312-326.

Dixit, A., and J.E. Stiglitz (1977), "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, 67, 297-308.

Gasmi, F., and J.J. Laffont and Q. Vuong (1992), "Econometric Analysis of Collusive Behavior in a Soft-Drink Market," *Journal of Economics & Strategy*, 1(2), 277-311.

Goldberg, P. (1995), "Product Differentiation and Oligopoly in International Markets: The Case of the Automobile Industry," *Econometrica*, 63, 891-951.

Gorman, W.M. (1959), "Separable Utility and Aggregation," *Econometrica*, 27, 469-81.

Gweke, J. (1995), "Monte Carlo Simulation and Numerical Integration," Federal Reserve Bank of Minneapolis, Research Department Staff Report 192.

Hansen, L.P., J. Heaton, and A. Yaron (1996), "Finite Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14(3), 262-80.

Hausman, J. (1996), "Valuation of New Goods Under Perfect and Imperfect Competition,"in T. Bresnahan and R. Gordon, eds., *The Economics of New Goods*, Studies in Income and Wealth Vol. 58, Chicago: National Bureau of Economic Research.

Hausman, J., G. Leonard, and J.D. Zona (1994), "Competitive Analysis with Differentiated

Products," *Annales D'Economie et de Statistique*, 34, 159-80.

Hausman, J., and D. Wise (1978), "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 49, 403-26.

McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, eds., *Frontiers of Econometrics*, New York, Academic Press.

McFadden, D. (1978), "Modeling the Choice of Residential Location," in A. Karlgvist, et al., eds., *Spatial Interaction Theory and Planning Models*, Amsterdam: North-Holland.

McFadden, D. (1984), "Econometric Analysis of Qualitative Response Models," in Z. Griliches and M. Intilligator, eds., *Handbook of Econometrics, Volume III*, Amsterdam: North-Holland.

McFadden, D. and K. Train (1997), "Mixed MNL Models for Discrete Response," University of California at Berkeley, mimeo.

Nelder, J.A., and R. Mead (1965), "A Simplex Method for Function Minimization," *Computer Journal*, Vol. 7, 308-313.

Nevo, A. (1997a), *Demand for Ready-to-Eat Cereal and Its Implications for Price Competition, Merger Analysis, and Valuation of New Goods*, Ph.D. Dissertation, Harvard University.

Nevo, A. (1997b), "Measuring Market Power in the Ready-to-Eat Cereal Industry," University of California at Berkeley, mimeo.

Nevo, A. (1997c), "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," University of California at Berkeley, mimeo.

Nevo, A. (1997c), "A Penny for Your Oats: The Value of a New Brand of Cereal," University of California at Berkeley, mimeo.

Pakes, A. (1986), "Patents as Options: Some Estimates of the Value of Holding A European Patent Stocks," *Econometrica*, 54, 755-784.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery (1994), *Numerical Recipes in C*, Cambridge University Press.

Rossi, P., R.E. McCulloch, and G.M. Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15(4), 321-40.

Simon, C.P. , and L. Blume (1994), *Mathematics for Economists*, New York: W. W. Norton & Company.

Spence, M. (1976), "Product Selection, Fixed Costs, and Monopolistic Competition," *Review of Economic Studies*, 43, 217-235.

Stone, J. (1954), "Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand," *Economic Journal*, 64, 511-527.

Theil, H. (1965), "The Information Approach to Demand Analysis," *Econometrica*, 6, 375-80.