Advanced Econometric Topics
ECO-R003 - Fall 2015

School *of* Economics
University of East Anglia

Demand Estimation
Lecture Notes

Lectures:	3pm-5pm (in JSC 1.01)	Instructor:	Farasat A.S. Bokhari
Internet:	http://www.uea.ac.uk/economics/	Email:	f.bokhari@uea.ac.uk
Office:	Arts 3.34	Tel:	+44 (160) 359-7534
Hours:	Wednesdays 2-4pm	Fax:	+44 (160) 345-6259

Lecture Notes on Demand Estimation

Last updated in December, 2013

Farasat A.S. Bokhari ©

f.bokhari@uea.ac.uk

1. Preliminaries

1.1. Why Demand Estimation?

Demand systems often form the bedrock upon which empirical work in industrial organization rests.[1] IO theory is mostly concerned about the supply-side. However, costs are important determinants of firm behavior which are usually unobserved. For instance, a fundamental issue is to measure market power, which is measured by the price-cost margin, or the Lerner index, as

$$\frac{p - mc}{p}. \tag{1.1}$$

The problem is that we do not always have data on marginal costs. The "new empirical industrial organization" (NEIO) literature is motivated by this data problem, where firms' mark-ups are obtained by estimating demand functions, and from these, mark-ups are then backed out. The intuition of this approach is most easily seen in the monopoly case. Consider the monopolist's maximization problem,

$$\max_p pq(p) - c(q(p)) \tag{1.2}$$

where $q(p)$ is the demand curve faced by the firm, and the first order conditions (FOC) imply

$$q(p) + p\frac{\partial q(p)}{\partial p} = \frac{\partial c(q(p))}{\partial q}\frac{\partial q(p)}{\partial p} = mc(q(p))\frac{\partial q(p)}{\partial p}. \tag{1.3}$$

At the optimal price, we get the relationship,

$$(p^* - mc(q(p^*))) = -\left.\frac{q(p)}{\partial q(p)/\partial p}\right|_{p=p^*} \tag{1.4}$$

or equivalently,

$$\frac{p^* - mc(q(p^*))}{p^*} = -\frac{1}{\eta(p^*)} \tag{1.5}$$

where $\eta(p^*) = \frac{p}{q(p)} \left.\frac{\partial q(p)}{\partial p}\right|_{p=p^*}$ is the price elasticity of demand. Thus, if can get a good estimate of elasticity, we can infer the markup.

[1] **Acknowledgements:** These lecture notes are based on a number of sources and draw heavily from the following articles/chapters: Cameron and Trivedi (2005, Chap. 6); Deaton and Muellbauer (1980, Chap. 3 & 5); Hausman et al. (1994); Bokhari and Fournier (2013); Berry (1994); Berry et al. (1995); Ackerberg et al. (2007); Nevo (2000, 2001). In addition to these primary sources, I have also benefitted from presentations/lecture notes on the same topics by other researchers who have generously put their slides on the internet. These sources include (1) Matthew Shum (Lecture notes: Demand in differentiated-product markets); (2) Matthijs Wildenbeest (Structural Econometric Modeling in Industrial Organization); (3) Eric Rasmusen (The BLP Method of Demand Curve Estimation in Industrial Organization); (4) John Asker and Allan Collard-Wexler (Demand Systems for Empirical Work in IO); (5) Jonathan Levin (Differentiated Products Demand Systems); (6) Ariel Pakes (NBERMetrics) and; (7) Aviv Nevo (NBER Methods Lecture – Estimation of Static Discrete Choice Models Using Market Level Data). Finally, I am also in debt to my colleague Franco Mariuzzo for providing significant feedback on these notes. All errors are mine.

This reasoning extends to oligopoly as well (covered later). Briefly though, when there are differentiated products, we want to estimate the system of demand equations and infer the markups using the full cross-elasticity matrix. These estimates can then be used in a variety of different contexts, including merger simulations to predict post merger prices, estimating the value of a new goods (via changes in consumer surplus), or other policy questions such as allowing direct to consumer advertising or parallel trade for pharmaceutical products etc. The process thus begins with estimating a system of demand equations.

Earlier empirical work focused on specifying representative consumer demand systems such that they allowed for various substitution patterns, and were consistent with economic theory. These methods included estimating the Linear Expenditure model (Stone, 1954), the Rotterdam model (Theil, 1965; and Barten 1966), or the more flexible ones such as the Translog model (Christensen, Jorgenson, and Lau, 1975), and the Almost Ideal Demand System (Deaton and Muellbauer, 1980a). In these lecture notes, we will discuss details of the AIDS model but within the context of multistage budgeting as well as variants of logit models derived from random utility/discrete choice models.

1.2. **Estimation issues and approaches to demand estimation**

Some common problems in demand estimation include endogeneity, multicollinearity, the dimensionality problem, and accounting for observed and unobserved heterogeneity among consumers. Depending on the context and the question, a researcher needs to be careful about choosing the appropriate estimation methodology, as there are tradeoffs between how well different methods deal with these issues and how relevant any given problem is with a context. The usual topology of various approaches to demand estimation is along the following lines:

- Single vs Multi-products
- Product or Characteristics Space approach to estimation
- Representative vs Heterogenous Agents

Single vs. Multiproduct Systems. Most products have substitutes or complements and it is often necessary to explicitly account for the substitution possibilities to adequately answer the research question at hand. In the context of multiproducts, the researcher also has to face the problem of dimensionality and multicollinearity. Consider a system of demand equations

$$\mathbf{q} = D(\mathbf{p}, \mathbf{z}; \boldsymbol{\theta}, \boldsymbol{\xi}), \tag{1.6}$$

where $\mathbf{q}$ is a $J \times 1$ vector of quantities, $\mathbf{p}$ is a vector of prices, $\mathbf{z}$ is a vector of exogenous variables that shift demand, $\boldsymbol{\theta}$ are the parameters to be estimated, and $\boldsymbol{\xi}$ are the error terms. In a system with J products, even with some simple and restrictive forms, the number of parameters to estimate is large. If for instance, $D(\cdot)$ is linear so that $D(\mathbf{p}) = \mathbf{Ap}$ where $\mathbf{A}$ is a $J \times J$ matrix of slope coefficients, then there are J^2 parameters to estimate (plus additional ones due to the exogenous variables $\mathbf{z}$). Imposing the symmetry of the Slutsky matrix or adding up restrictions (Engle and Cournout aggregation) reduces the number of parameters to be estimated, however, the essential problem, that the number of parameters increases in the square of the number of products, remains.[2] This problem of too many parameters is augmented when we attempt to estimate flexible demand systems.

Further, as prices of related products often move together, lack of precision of estimated parameters due to multicollinearity is another typical problem. Similarly, when $\text{cov}(p, \xi) \neq 0$, so that the prices are endogenous, the econometrician has to find at least as many instruments as the number of products. It is usually difficult to find instruments that are both exogenous and will not generate moment conditions that are nearly collinear.

One way of avoiding the problem of estimating too many parameters is to start with more restrictive forms. Consider starting with a constant elasticity of substitution (CES) utility function and then deriving a demand system from it. For J different products, the CES utility function takes the form

$$u(\mathbf{q}; \rho) = u(q_1, q_2, \ldots, q_J; \rho) = \Big(\sum_i^J q_i^\rho \Big)^{1/\rho} \tag{1.7}$$

where ρ is the parameter of interest that measures the elasticity of substitution. The demand for a representative consumer is then given by,

$$q_j(\mathbf{p}, I; \rho) = \frac{p_j^{1/(1-\rho)}}{\sum_i^J p_i^{\rho/(1-\rho)}} I \qquad j = 1, \ldots, J. \tag{1.8}$$

In the system above, I is the income of the representative consumer, and the dimensionality problem is solved by imposing symmetry between different products, i.e. all the cross-price elasticities are now the same. Specifically, the cross elasticity between products i and j is the same as between k and j for all combinations of i, j, k,

$$\frac{\partial q_i}{\partial p_j} \frac{p_j}{q_i} = \frac{\partial q_k}{\partial p_j} \frac{p_j}{q_k} \qquad \forall i, j, k. \tag{1.9}$$

Thus, the while the dimensionality problem is solved – instead of estimating J^2 parameters, we need to estimate only a single parameter – it comes at the high cost of imposing such a restrictive

[2] The Slutsky equation decomposes the change in demand for good j as a response to a change in the price of good i as $\frac{\partial q_j}{\partial p_i} = \frac{\partial h_j}{\partial p_i} - q_i \frac{\partial q_j}{\partial y}$ where q_j and h_j are the marshallian and hicksian demand functions respectively for product j, and y is the income or total expenditure, and the relation holds for all $i, j = 1, \ldots, J$ and $i \neq j$. The adding up restrictions are the Engle aggregation ($\sum_j s_j \eta_{jy} = 1$) and Cournot aggregation ($\sum_j s_j \eta_{ji} = -s_i$) where η_{ji} is the cross price elasticity of product j with respect to price of i, η_{jy} is the income elasticity of product j and s_i, s_j are the expenditure shares.

substitution pattern which may not be very realistic as some products may be much closer in attributes to each other than other products and their cross price elasticities may be much higher.

An alternative to the single parameter of the CES utility function is the logit demand (Anderson, de Palma, and Thisse, 1992). This has a richer substitution pattern, and is derived from

$$u(\mathbf{q};\boldsymbol{\delta}) = \sum_{j}^{J} \delta_j q_j - \sum_{j}^{J} q_j \ln q_j. \tag{1.10}$$

Elasticities in this model depend on market shares (given by J number of parameters δ_j) but again not on the similarities among the products.

Other ways researchers sometime get around this dimensionality problem is by imposing more structure in the form of functional form assumptions on utility function – separability and multistage budgeting – which leads to "grouping" or "nesting" approaches. In this case, we group products together and consider substitution across and within groups as separate things. This however requires some ex-ante assumptions about how a consumer chooses a particular product. We consider this approach in more detail later on.

Endogeneity. To see why prices can be endogenous, consider a simple linear demand/supply model for a single homogenous product over T markets, where aggregate demand/supply relations are given by

$$\begin{aligned} q_t^d &= \beta_{10} + \gamma_{12} p_t + \beta_{11} x_{1t} + \xi_{1t}, \\ p_t &= \beta_{20} + \gamma_{22} q_t^s + \beta_{22} x_{2t} + \xi_{2t}, \\ q_t^s &= q_t^d. \end{aligned} \tag{1.11}$$

In the equations above, the error terms are such that[3]

$$\begin{aligned} E(\xi_{1t}|\mathbf{x}_t) = 0, E(\xi_{2t}|\mathbf{x}_t) = 0, \\ E(\xi_{1t}^2|\mathbf{x}_t) = \sigma_1^2, E(\xi_{2t}^2|\mathbf{x}_t) = \sigma_2^2 \\ E(\xi_{1t}\mathbf{x}_t) = 0, E(\xi_{2t}\mathbf{x}_t) = 0, \\ \text{and } E(\xi_{1t}\xi_{2t}|\mathbf{x}_t) = 0 \end{aligned} \tag{1.12}$$

where $\mathbf{x}_t = [1\ x_{1t}\ x_{2t}]$. Thus, the error terms in the two equations are mean zero with variances σ_1^2, σ_2^2, the x_1 and x_2 are exogenous variables (demand and supply shifters) and suppose that *the error terms across the two equations are uncorrelated* (i.e., $E(\xi_{1t}\xi_{2t}|\mathbf{x}_t) = 0$). To see how endogeneity

[3]Since we have already made the stronger assumption that $E(\xi_{1t}|\mathbf{x}_t) = 0$, technically we do not need to explicitly make the assumption that $E(\xi_{1t}\mathbf{x}_t) = 0$, since the latter is implied by the former assumption of zero conditional mean due to law of iterated expectations. Nonetheless, I include it just to be clear.

arises, explicitly solve for the reduced form equilibrium values of q^* and p^*. If we solve the equations above (dropping the t subscript), we get

$$\begin{aligned} q^* &= \frac{\beta_{10}+\beta_{20}\gamma_{12}}{1-\gamma_{12}\gamma_{22}} + \frac{\beta_{11}}{1-\gamma_{12}\gamma_{22}}x_1 + \frac{\gamma_{12}\beta_{22}}{1-\gamma_{12}\gamma_{22}}x_2 + \frac{\xi_1+\gamma_{12}\xi_2}{1-\gamma_{12}\gamma_{22}} \\ p^* &= \frac{\beta_{20}+\beta_{10}\gamma_{22}}{1-\gamma_{12}\gamma_{22}} + \frac{\beta_{11}\gamma_{22}}{1-\gamma_{12}\gamma_{22}}x_1 + \frac{\beta_{22}}{1-\gamma_{12}\gamma_{22}}x_2 + \frac{\gamma_{22}\xi_1+\xi_2}{1-\gamma_{12}\gamma_{22}} \end{aligned} \tag{1.13}$$

which shows that p^* is a function of ξ_1 (and ξ_2) and hence an OLS estimation of the demand equation above (regress q on p, x_1) will result in an inconsistent estimate of γ_{12} and other parameters. Similar issue applies to OLS estimation of the supply equation. It is instructive to explicitly compute the conditional covariance (conditional on $\mathbf{x}_t$) between p and ξ_1 (or between q and ξ_2 for the supply equation). To do so, first note that conditional on $\mathbf{x}_t$,

$$\begin{aligned} p^* - E(p^*) &= \frac{\gamma_{22}\xi_1+\xi_2}{1-\gamma_{12}\gamma_{22}} \\ \text{and } \xi_1 - E(\xi_1) &= \xi_1. \end{aligned} \tag{1.14}$$

Thus, conditional on $\mathbf{x}_t$

$$\text{cov}(p,\xi_1) = \frac{\gamma_{22}}{1-\gamma_{12}\gamma_{22}}\sigma_1^2 + \frac{\mathrm{E}(\xi_1\xi_2)}{1-\gamma_{12}\gamma_{22}}. \tag{1.15}$$

Note that even if the error terms across the two equations were uncorrelated ($\mathrm{E}(\xi_{1t}\xi_{2t}|\mathbf{x}_t = 0)$, the covariance between p and ξ_1 would still not be zero. On the other hand, if γ_{22} is zero, q does not appear in the supply equation, i.e., it is a triangular system of equations and OLS estimation is fine as long as $\mathrm{E}(\xi_{1t}\xi_{2t}|\mathbf{x}_t) = 0$. Finally, for completeness – complete system of equations, i.e., the number of equations are equal to the number of endogenous variables – we also require that $\gamma_{12} \neq 1/\gamma_{22}$.

For future reference, we can re-write the system in (1.11) in matrix notation. Let

$$\mathbf{y}_t' = \begin{bmatrix} q_t & p_t \end{bmatrix}, \mathbf{x}_t = \begin{bmatrix} 1 & x_{1t} & x_{2t} \end{bmatrix}, \boldsymbol{\xi}_t' = \begin{bmatrix} \xi_{1t} & \xi_{2t} \end{bmatrix}, \boldsymbol{\Gamma} = \begin{bmatrix} 1 & -\gamma_{22} \\ -\gamma_{12} & 1 \end{bmatrix} \text{ and, } \mathbf{B} = \begin{bmatrix} \beta_{10} & \beta_{20} \\ \beta_{11} & 0 \\ 0 & \beta_{22} \end{bmatrix} \tag{1.16}$$

then, the system of equations above can be written as

$$\mathbf{y}_t'\boldsymbol{\Gamma} - \mathbf{x}_t\mathbf{B} = \boldsymbol{\xi}_t', \tag{1.17}$$

so that the reduced form equation is

$$\mathbf{y}_t' = \mathbf{x}_t\boldsymbol{\Pi} + \mathbf{v}_t' \qquad \text{where } \boldsymbol{\Pi} = \mathbf{B}\boldsymbol{\Gamma}^{-1} \text{ and, } \mathbf{v}_t' = \boldsymbol{\xi}_t'\boldsymbol{\Gamma}^{-1}. \tag{1.18}$$

Note that in the equation above we are takeing the inverse of the $\boldsymbol{\Gamma}$. But the inverse exists if the determinant ($\det(\boldsymbol{\Gamma}) = 1 - \gamma_{12}\gamma_{22}$) is not zero, which goes back to the condition $\gamma_{12} \neq 1/\gamma_{22}$ mentioned above. The moment restrictions in (1.12) (in general we do not need to impose $E(\xi_{1t}\xi_{2t}|\mathbf{x}_t) = 0$) are

$$\begin{aligned} &\mathrm{E}(\boldsymbol{\xi}_t|\mathbf{x}_t) = \mathbf{0}, && \mathrm{E}(\boldsymbol{\xi}_t\boldsymbol{\xi}_t'|\mathbf{x}_t) = \boldsymbol{\Sigma} \\ &\mathrm{E}(\mathbf{v}_t|\mathbf{x}_t) = \mathbf{0}, && \mathrm{E}(\mathbf{v}_t\mathbf{v}_t'|\mathbf{x}_t) = \boldsymbol{\Omega} \\ &\text{where } \boldsymbol{\Omega} = (\boldsymbol{\Gamma}^{-1})'\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1}. \end{aligned} \tag{1.19}$$

In this case, estimation can proceed with IV/2SLS (or 3SLS for joint estimation), where the demand equation is estimated using x_{2t} as the instrument, and supply equation is estimated using x_{1t} as the instrument. If either $\beta_{22} = 0$ or if data on x_{2t} is not available, demand equation cannot be identified/estimated consistently (vice versa for supply equation). Since the x's are exogenous variables, they can serve as instruments.

- x_{2t} are cost shifters. They affect production costs. Correlated with p_t but not with ξ_{1t}: use as instruments in demand function.
- x_{1t} are demand shifters. Affect willingness-to-pay, but not a firm's production costs. Correlated with q_t but not with ξ_{2t}: use as instruments in supply function.

Instrument: The broadest definition of an instrument is a variable z such that for all possible values of z:

$$\Pr[z|\xi] = \Pr[z|\xi']. \tag{1.20}$$

But for certain values of X we have

$$\Pr[x|z] = \Pr[x|z']. \tag{1.21}$$

So the intuition is the that z is not affected by ξ, but has some effect on x. The usual way to express these conditions is that an instrument is such that: $E[z\xi] = 0$ and $E[xz] \neq 0$.

Product vs Characteristics Space. There are two general approaches to estimating demand. Product space is more natural in the sense that consumers have preferences over products, and those preferences lead to demand at the product level. The alternative, the characteristics space approach (Lancaster, 1966; McFadden, 1973), views products as bundles of characteristics and consumers have preferences over those characteristics. Thus each individual's demand for a given product is just a function of the characteristics of the product. To be specific, we can think of a set of products (Toyota Minivan, Lexus SUV, etc.) or we can think of them as a collection of various properties (horsepower, size, color, etc.). In general, demand systems in characteristic space are approximations to product space demand systems and hence, we can either model consumers as having preferences over products, or over characteristics (note that not all of the characteristics need to be observed and may form part of the error term). This changes the "space" in which we analyze the demand system. When working in characteristics space, a researcher begins by specifying a utility function over the characteristics of the product and then derives observable market demand from it. We will study this in more detail later in the context of discrete choice models. The two approaches to demand estimation have tradeoffs in terms of relative strengths and weaknesses in terms of how they deal with different issues, some of which are discussed below.

- For large number of products (say $J = 50$), the product space approach leads to the dimensionality problem mentioned earlier, and may require grouping/nesting these products. By contrast, if we can reduce J products to just a few K characteristics, and the preferences over those characteristics are, say normally distributed, then we have to estimate K means and $K(K+1)/2$ covariances. *If* there were no unobserved characteristics, then $K(1+(K+1)/2)$ parameters would suffice to analyze own and cross price elasticities for all J goods.
- If there are too many characteristics (K is large), then the problem of too many parameters re-appears as in the product space, and we need data on each of these characteristics. A solution is to model some of them as unobserved characteristics – but this leads to the endogeneity problem if the unobserved characteristics (think product quality) are correlated with the price, which they usually are.
- A related problem is that of weak instruments in the product space approach. Consider the case of demand equation for the quantity of the j^{th} product. In product space, it would be a function of all other prices, so that $q_j = f(p_1, \ldots, p_J)$ and we have J such equations, one for each product. The estimation is thus in 'wide form' and we need to find instruments for each of these prices. In the characteristics space approach, the fundamental issue is the same, but often the resulting functional form is such that estimation is done in 'long form', i.e., the left hand side of the equation, which is typically shares, is also a function of prices, but appears in the equation as a single variable $s_j = f(p_j)$ and there are J such rows per market. Thus, in the product space, when we do a 2SLS estimation, where the first stage involves regressing each price variable on the exogenous set of instruments, we effectively have to estimate J number of first-stage regressions. By contrast, in the characteristics space, we have to do only one first-stage regression as data on all prices is a single column vector that is regressed on a column vector of instruments. Effectively then, the first stage regressions can exhibit weak instruments property in the product space set up while they may not exhibit such a property in the characteristics space approach.
- If we are interested in the counterfactual exercise to assess the welfare impact of a new introduction in an ex-ante period (say a new proposed generic drug or a me-too drug), it is difficult to do so in the product space (we can do it using ex-post data though), but it is easier to do this exercise using the characteristic space approach. This is because if we have estimated the demand system using the characteristic approach, and we know the proposed

characteristics of the new good, we can, in principle, analyze what the demand for the new good would be. Note however that if the new good is totally different from products already in the market, i.e., has very different (and new) properties, characteristics space approach may not help either (e.g., could we have predicted the demand for laptops based on the characteristics of desktop computers, or for a new drug which proposes treatment of a formally un-treatable disease?)

- Most of the characteristics space estimation, at least on aggregate data, does not easily lend to analyzing products which are used in bundles or as complements. This is an on-going area of research.

Representative or Heterogenous Consumer. Consider the demand function of single product j in market t for a representative consumer, given by

$$q_{jt} = \gamma_j + \alpha_j p_{jt} + \mathbf{x}_{jt}\beta_j + \xi_{jt}, \tag{1.22}$$

where $\mathbf{x}_{jt}$ is a vector of product characteristics and ξ_{jt} are the unobserved components of demand. The interest is in estimating α_j and demand elasticity (for instance, if the demand function is isoelastic so that the left hand side of the equation above is $\ln(q_{jt})$, then the elasticity is $\epsilon_{jjt} = \alpha_j p_{jt}$). Even though product specific intercepts γ_j (willingness to pay) have been included in the model, they are demand shifters, and as such do not change the sensitivity to price depending on the level of income or other demographic characteristics such as family size. However, micro studies show that the price coefficient depends in an important way on income/wealth, i.e., lower income people care more about price. Consequently, if the income distribution varies across the markets, we should expect the price coefficient to vary across these markets, and we need to find a way to allow for it. One could make γ_j to be a function of income, but they are still demand shifters and do not change the sensitivity to price. Similarly, other demographic differences may be important to model as well. Such a representative consumer model does not allow sensitivity to price to vary by income. One could potentially include some ad-hoc interaction terms between average values of demographic variables in market t with price (and other product characteristics) but unless such interactive terms can be derived from the utility function, the equations being estimated may not represent demand derived from a consumer's utility maximization problem. To make it a heterogenous agent model, it is more typical to build a micro model where the parameters that enter the utility function of a consumer – say γ_j and α_j – vary over individuals and are perhaps functions of their demographics. In that case, the demand equations to be estimated would end up looking something like

$$q_{jt} = \int \gamma_{ij} dG(\gamma_{ij}) + \int \alpha_{ij} p_{jt} dF(\alpha_{ij}) + \mathbf{x}_{jt}\beta_j + \xi_{jt} \tag{1.23}$$

where γ_{ij} and α_{ij} are person and product specific random intercepts and slope coefficients, with known or assumed distribution functions $\gamma_{ij} \sim G(\gamma|\tau)$ and $\alpha_{ij} \sim F(\alpha|\theta)$, and where θ and τ are parameters to be estimated and are functions of demographic variables. This is called a random coefficients model and we will take up this approach within the context of discrete choice modeling.

2. Product Space Approaches

The main method we will look at in the products space approach is one which solves the dimensionality problem by dividing the products into small sub-groups and then allowing some relatively flexible substitution patterns between the products within a group. To this end, it would be useful if we could break down the overall consumer decision problem into separate parts, some of which could be estimated separately. This is the issue of **separability** in demand theory. A related problem is that of **aggregation**, which considers the relationship between individual consumers' behavior and aggregate consumer behavior (which is the sum of individual behavior over all individuals). This is a longstanding issue in demand theory, i.e., how to relate the demand system of a group of consumers to the underlying demand of individual consumers. There is no reason why aggregate data, or any data that are an average over many people should conform to a theory of consumer behavior that focuses on individual people or households. Nonetheless, study of aggregate data may allow us to say something about individual behavior. Thus, when working with aggregate data, one can ask whether there are assumptions on preferences such that aggregate demand is generated by a "representative consumer" with "rationalizeable" preferences. We start with the latter issue.

2.1. Homothecity, Gorman Polar Form and Aggregation

Homothecity. We start with a very simple assumption on preferences, that they are **homothetic**, and see what it may imply about demand curves. Preferences ($\succeq$) are homothetic if $t\mathbf{q}_1 \succeq t\mathbf{q}_2 \Leftrightarrow \mathbf{q}_1 \succeq \mathbf{q}_2$ for any $t > 0$ i.e., the consumer is indifferent between bundles $t\mathbf{q}_1$ and $t\mathbf{q}_2$ whenever they are indifferent between bundles $\mathbf{q}_1$ and $\mathbf{q}_2$. Thus, with homothetic preferences, there is only one indifference curve and any indifference curve is a radial blowup of another, and all indifference sets are related by proportional expansion along rays. The implication is that marginal rates of substitution are unaffected by equal proportional changes in all quantities, so that income expansion paths are straight lines through the origin. In terms of utility functions, preferences are homothetic if and only if they are of the form

$$u(\mathbf{q}) = F(f(\mathbf{q})) \text{ where} f(\cdot) \text{ is a function such that } f(t\mathbf{q}) = tf(\mathbf{q}), \tag{2.1}$$

and $F(\cdot)$ is a monotone increasing function. Thus, the utility function must admit a function that is homogenous of degree one (the $f(\cdot)$) and since utility functions are only defined upto monotonic transformations, then we may as well write the utility function to be just $u(\mathbf{q}) = f(\mathbf{q})$ where the latter is, as before, homogeneous of degree one.

Consider the consumer's expenditure minimization problem $\min \mathbf{p} \cdot \mathbf{q}$ s.t. $u(\mathbf{q}) = f(\mathbf{q}) = u$. Since the function is homogenous of degree one, doubling $\mathbf{q}$ will double the target utility, but doubling $\mathbf{q}$ means doubling the expenditure. This means that if $e(\mathbf{p}, u) = \mathbf{q}^* \cdot \mathbf{p}$ is the minimum expenditure for target utility u, then for a target utility of tu, the minimum expenditure is $e(\mathbf{p}, tu) = t\mathbf{q}^* \cdot \mathbf{p} = te(\mathbf{p}, u)$. Now If the initial target utility is equal to 1, then by letting $t = u$, we can write $e(\mathbf{p}, u) = ue(\mathbf{p}, 1)$ and hence, for homothetic utility preferences, the expenditure function is of the form

$$e(\mathbf{p}, u) = ub(\mathbf{p}), \tag{2.2}$$

where $b(\mathbf{p})$ is some linearly homogenous and concave function of prices. Similarly, in terms of indirect utility $V(\mathbf{p}, y)$ and demand functions (hicksian, $h(\mathbf{p}, u)$ and marshallian $q(\mathbf{p}, y)$), a homothetic utility function implies the following forms

$$\begin{aligned} &V(\mathbf{p}, y) = \frac{y}{b(\mathbf{p})}, \qquad q_j^h(\mathbf{p}, u) = u\frac{\partial b(\mathbf{p})}{\partial p_j}, \qquad q_j(\mathbf{p}, y) = yq_j(\mathbf{p}), \\ &\text{where } y = \sum_j p_j q_j \text{ is the total expenditure.} \end{aligned} \tag{2.3}$$

A leading example of homothetic preference is the cobb-douglas utility function given by $u(\mathbf{q}) = q_1^{\beta_1} q_2^{\beta_2} \ldots, q_J^{\beta_J}$ where the associated demand functions are of the form

$$q_j = y\frac{1}{p_j}\frac{\beta_i}{\sum_j^J \beta_j}.$$

The demand for each good is proportional to expenditure (income), or alternatively, the Engel curve for each good is a straight line going through the origin. To see this more clearly, take the log of $q_j(\mathbf{p}, y)$ in (2.3), so that $\ln q_j = lny + \ln q_j(\mathbf{p})$, we see that the expenditure elasticity of good j is always one

$$\eta_j = \frac{\partial lnq_j}{\partial \ln y} = 1 \qquad \forall j = 1, \ldots, J.$$

This is known as the expenditure proportionality, which is equivalent to the requirement that budget shares ($w_j = \frac{p_j q_j}{y}$) of all commodities are independent of the level of total expenditure (income) so that a consumer always spends a constant proportion of their income on a product, even though income may be varying across different consumers. Thus, with homothetic preferences, all expenditure elasticities are equal to one – a result that is contradicted by most empirical studies. A result that follows is that with identical homothetic preferences, aggregate demand is "as if" there were a single consumer with the same preferences and the total income of all consumers. Note also that demand for each good is independent of prices of other products implying that cross-price elasticities are zero.

Quasi-Homothecity and Gorman Polar Form. A less restrictive form is that of **quasi-homotheticity**. In this formulation, a fixed expenditure element ($a(\mathbf{p})$) is added to the expenditure function in equation (2.2) so that it is now given by

$$e(\mathbf{p}, u) = a(\mathbf{p}) + ub(\mathbf{p}). \tag{2.4}$$

This form is called the **Gorman Polar Form**. The term $a(\mathbf{p})$ represents the subsistence level of expenditure when $u = 0$ and $b(\mathbf{p})$ is the marginal cost of utility. For the expenditure function to be concave in prices it is necessary and sufficient for $a(\mathbf{p})$ and $b(\mathbf{p})$ to be concave in prices. The associated indirect utility and demand functions (per the usual derivations) take the forms

$$\begin{aligned} &V(\mathbf{p}, y) = \frac{y - a(\mathbf{p})}{b(\mathbf{p})} \quad \text{and} \quad q_j(\mathbf{p}, y) = a_j(\mathbf{p}) + \frac{b_j(\mathbf{p})}{b(\mathbf{p})}\Big[y - a(\mathbf{p})\Big] \\ &\text{where} \\ &a_j(\mathbf{p}) = \frac{\partial a(\mathbf{p})}{\partial p_j} \quad \text{and} \quad b_j(\mathbf{p}) = \frac{\partial b(\mathbf{p})}{\partial p_j}. \end{aligned} \tag{2.5}$$

For the indirect utility function above, $a(\mathbf{p})$ is interpreted as the subsistence spending amount and $b(\mathbf{p})$ is a price index that deflates income/expenditure over and above the subsistence level. Before moving on, it is worth rewriting (2.5) in an alternative form, as it is sometimes useful and because we will later use the alternative form. Note that we can define $A(\mathbf{p}) = \frac{1}{b(\mathbf{p})}$ and $B(\mathbf{p}) = -\frac{a(\mathbf{p})}{b(\mathbf{p})}$ in which case the indirect utility can be written as $V(\mathbf{p}, y) = A(\mathbf{p})y + B(\mathbf{p})$. In this case $q_j(\mathbf{p}, y) = a_j(\mathbf{p}) + b_j(\mathbf{p})B(\mathbf{p}) + b_j(\mathbf{p})A(\mathbf{p})y$. Now if we further define $\alpha_j(\mathbf{p}) = a_j(\mathbf{p}) + b_j(\mathbf{p})B(\mathbf{p}) = a_j(\mathbf{p}) - \beta_j(\mathbf{p})a(\mathbf{p})$ and $\beta_j(\mathbf{p}) = b_j(\mathbf{p})A(\mathbf{p}) = \frac{b_j(\mathbf{p})}{b(\mathbf{p})}$, then (2.4) and (2.5) can be expressed as

$$\begin{aligned} &e(\mathbf{p}, u) = a(\mathbf{p}) + ub(\mathbf{p}) \\ &V(\mathbf{p}, y) = A(\mathbf{p})y + B(\mathbf{p}) \\ &q_j(\mathbf{p}, y) = \alpha_j(\mathbf{p}) + \beta_j(\mathbf{p})y, \text{ where,} \\ &A(\mathbf{p}) = \frac{1}{b(\mathbf{p})}, B(\mathbf{p}) = -\frac{a(\mathbf{p})}{b(\mathbf{p})} \text{ and,} \\ &\alpha_j(\mathbf{p}) = \frac{\partial a(\mathbf{p})}{\partial p_j} - \beta_j(\mathbf{p})a(\mathbf{p}) \text{ and, } \beta_j(\mathbf{p}) = \frac{1}{b(\mathbf{p})}\frac{\partial b(\mathbf{p})}{\partial p_j}. \end{aligned} \tag{2.6}$$

The budget share equations in this case are given by a weighted average of two terms

$$w_j = \Big(\frac{a}{y}\Big)(\frac{p_j a_j}{a}) + \Big(1 - \frac{a}{y}\Big)(\frac{p_j b_j}{b}), \tag{2.7}$$

where if $a = y$ (subsistence level is equal to the entire income) the budget share of good j is equal to just $\frac{p_j a_j}{a}$, and if expenditure is much larger than the subsistence level (so $a/y \approx 0$) then the share is given by $\frac{p_j b_j}{b}$. In aggregate, the expenditure patterns are a weighted average of value shares appropriate to very rich and very poor consumers. As with homothetic preferences, Engle curves are still linear but they do not go through the origin anymore. Consequently, although homotheticity implies unitary income elasticities for all commodities, quasi-hometheticity implies elasticities that only tend to unity as total expenditure increases. This is a significant generalization/improvement over the previous case, but still restrictive as it unlikely to be true for narrowly defined commodities. Even for broad commodities such as food, household budget studies tend to give nonlinear Engel curves (we will get to that further below).

A leading example of quasi-homothetic preferences is the Stone-Geary utility function given by $u(\mathbf{q}) = \prod_j^J (q_j - \alpha_j)^{\beta_j}$ or equivalently as $u(\mathbf{q}) = \sum_j^J \beta_j \ln(q_j - \alpha_j)$ with $\sum_j^J \beta_j = 1$ where, as before, the α_j are the subsistence levels for each good. For this utility function, the implied expenditure, indirect utility and demand functions are

$$e(\mathbf{p}, u) = \sum_i^J p_j\alpha_j + u\prod_j^J p_j^{\beta_j}, \qquad V(\mathbf{p}, y) = \frac{y - \sum_j^J p_j\alpha_j}{\prod_j^J p_j^{\beta_j}},$$

$$\text{and} \qquad q_j(\mathbf{p}, y) = \alpha_j + \beta_j \frac{y - \sum_j^J p_j\alpha_j}{p_j}$$

which are of the same general forms as discussed above (i.e, $a(\mathbf{p}) = \sum_i^J p_j\alpha_j$ and $b(\mathbf{p}) = \prod_j^J p_j^{\beta_j}$). Starting with the expenditure function, $a(\mathbf{p}) = \sum_i^J p_j\alpha_j$ is the fixed expenditure for good j (with no substitution) plus a term that allows utility to be bought at a constant price per unit ($b(\mathbf{p}) = \prod_j^J p_j^{\beta_j}$). Since $\beta's$ add up to one, the later term can be thought of as a weighted geometric mean of the prices, and hence a price index representing the marginal cost of living. Similarly, in the indirect utility function, we get the interpretation of 'real' expenditure: since $\alpha's$ are the subsistence level, the discretionary levels $(y - \sum_j^J p_j\alpha_j)$ are deflated by the price index to give a real measure of welfare. Finally, note that in the demand function, the consumer obtains the subsistence level α_j of product j and the residual income $(y - \sum_j^J p_j\alpha_j)$ is allocated between different goods in fixed proportions β_j. The parameters β_j are called the marginal budget shares. Finally, note that the expenditure on good j is simply

$$p_jq_j = p_j\alpha_j + \beta_j(y - \sum_j^J p_j\alpha_j)$$

and is called the **linear expenditure system** (LES) (expenditure is linear in prices and income) which is easy to estimate, and has been very popular in empirical studies for this reason. It is completely characterized by the marginal budget share and subsistence level parameters, requiring estimation of $2J$ parameters, (2J - 1) of which may be chosen independently: compare that to the more general case of estimating $J^2 + J$ parameters (own and cross-price elasticities and income/expenditure elasticities), or, if adding up, homogeneity, and symmetry restrictions are imposed, there are $(2J - 1)(J/2 + 1)$ parameters to be estimated. Nonetheless, LES comes with its own short comings. If concavity of the expenditure function is allowed, then by construction all cross price elasticities are positive and hence the system cannot be used if some of the products are complements. Additionally, it turns out that there is an approximate proportionality between own price and expenditure elasticities.

Aggregation. Aggregate demand data raises the problem as to whether the aggregate demand function is consistent with consumer theory. This problem is referred to as the "aggregation problem". To overcome the aggregation problem, certain conditions are necessary under which we can treat aggregate demand estimations as resulting from the behavior of a single utility maximizing consumer (exact aggregation). Suppose there are N consumers (or households) that face the same prices but differ only in the incomes or expenditures on different products so that the demand for good j for the n^{th} individual is of form

$$q_{in} = g_{jn}(\mathbf{p}, y_n). \tag{2.8}$$

Then the average demand $\bar{q}_j$ – aggregated by adding up *quantities* over all individuals and dividing by N – is given by some function f_j as

$$\bar{q}_j = f_j(\mathbf{p}, y_1, y_2, \ldots y_N) = \frac{1}{N}\sum_{n}^{N} g_{jn}(\mathbf{p}, y_n). \tag{2.9}$$

Exact aggregation is possible if we can write (2.9) in the form

$$\bar{q}_j = g_j(\mathbf{p}, \bar{y}) \text{ where } \bar{y} = \frac{1}{N}\sum_{n}^{N} y_n. \tag{2.10}$$

Note that while (2.9) depends on the distribution of expenditures (incomes) $y_1, y_2, \ldots, y_N$, equation (2.10) does not depend on its distribution. This implies that for the equation to hold, any reallocation of income from one individual to another will not change the market demand. But this can only happen if the J different marginal propensities to spend are identical for all N consumers. Thus, both the high and low income individuals must allocate changes in income in exactly the same way. In turn, this implies that the general function in (2.8) must be linear in y_n, that is, for some function α_{jn} and β_j of $\mathbf{p}$ alone, be of form

$$q_{jn}(\mathbf{p}, y_n) = \alpha_{jn}(\mathbf{p}) + \beta_j(\mathbf{p})y_n. \tag{2.11}$$

Thus, if the aggregate (average) demand is a function of prices and average income, as in (2.10), then the underlying individual demand must be of form given by (2.11). But this is the same demand function from quasi-homothetic preferences as in (2.6) with a subscript n for the n^{th} consumer, and α_j and y both vary over consumers, but importantly, β_j does not vary over consumers (i.e, person specific $\alpha(\mathbf{p})$ but identical $\beta(\mathbf{p})$). Conversely, if the n^{th} consumer has quasi-homothetic preferences with demand given by (2.11), then the average demand – aggregated via adding up *quantities* over all individuals and dividing by N – is

$$\begin{aligned}\bar{q}_j &= \frac{1}{N}\sum_n^N q_{jn}(\mathbf{p}, y_n) \\ &= \alpha_j(\mathbf{p}) + \beta_j(\mathbf{p})\bar{y}, \text{ where} \\ \alpha_j(p) &= \frac{1}{N}\sum_n^N \alpha_{jn}(\mathbf{p}), \quad \text{and } \bar{y} = \frac{1}{N}\sum_n^N y_n.\end{aligned} \tag{2.12}$$

Thus, the aggregate demand is also quasi-homothetic and (2.11) is necessary and sufficient for (2.10). Note however that, the forms above are arising only due to aggregation requirements, and have nothing to do with requiring aggregate utility maximization. Suppose now that individuals maximize utility and the individuals demand function is of form (2.11). Gorman showed that quasi-homothetic demand of the form above is generated by consumer with the expenditure function given by

$$e_n(\mathbf{p}, u_n) = a_n(\mathbf{p}) + u_n b(\mathbf{p}), \tag{2.13}$$

i.e., expenditure is of (Gorman) polar form with subscript n in equation (2.6). Infact, Deaton and Muellbaur show that it is a 'if and only if' condition (see p. 151 and exercise 6.3). Similarly, the average of the expenditure functions in (2.13) is

$$\bar{e}(\mathbf{p}, u_n) = \bar{a}(\mathbf{p}) + ub(\mathbf{p}), \tag{2.14}$$

and corresponds to expenditure function for the average demand function in (2.12). Hence, if individuals maximize utility, and preferences are such that they satisfy the exact aggregation condition, then the average demand function will be consistent with utility maximization.

Nonlinear Aggregation. The aggregation given above leads to the linear Engel curves. Muellbauer (1975,1976) introduced exact nonlinear aggregation by starting with budget shares rather than with quantities, so that aggregation is over the budget shares of different consumers. The aggregate budget share of the j^{th} product $\bar{w}_j$, is defined as a weighted average of individual shares w_{jn} with weights given by the share of each individual in total expenditure on good j. Specifically, let the average budget share of good j be given by

$$\bar{w}_j = \frac{p_j \sum_n q_{jn}(\mathbf{p}, y_n)}{\sum_n y_n} = \sum_n \left(\frac{y_n}{\sum_n y_n}\right) w_{jn}. \tag{2.15}$$

In general, $\bar{w}_j$ is a function of prices and each individual's total expenditure/income. If we restrict $\bar{w}_j$ to be some function of prices and <u>average</u> expenditure $\bar{y}$, it leads back to linear aggregation case discussed above. Thus, Muellbauer instead makes $\bar{w}_j$ a function of prices and y_0, which is the total income/expenidture of a *representative* consumer and may itself be a function of the distribution of individual expenditures as well as of prices,

$$\sum_n \Big(\frac{y_n}{\sum_n y_n}\Big) w_{jn} = w_j(y_0(y_1, \ldots, y_J, \mathbf{p}), \mathbf{p}). \tag{2.16}$$

If the average share holds as the above given function, then the aggregate demand can be thought of as from a utility maximizing representative consumer with total income of y_0 and facing prices $\mathbf{p}$. Formally, the representative consumer exists if a indirect utility function $\psi(\mathbf{p}, y)$ and a corresponding expenditure function $e(\mathbf{p}, u)$ can be defined so that for some utility $u_0 = \psi(\mathbf{p}, y_0)$, we get

$$\bar{w}_j = w_j(\mathbf{p}, u_o) = \frac{\partial \ln e(\mathbf{p}, u_o)}{\partial \ln p_j} = \sum_n \Big(\frac{y_n}{\sum_n y_n}\Big) \frac{\partial \ln e_n(\mathbf{p}, u_n)}{\partial \ln p_j} \tag{2.17}$$

where the $e_n(\mathbf{p}, u_n)$ is the n^{th} consumers expenditure function with utility $u_n = \psi_n(\mathbf{p}, y_n)$. This representative budget share function shows that, although expenditure redistribution happens among the consumers, the representative consumer utility function, u_0, does not change. Therefore, this function can capture the change of u_n or different preferences among consumers while keeping u_0 constant. It turns out that such a representative consumer (and the assumed cost function) exists only if the preferences are such that the expenditure function of each individual has the form

$$e_n(\mathbf{p}, u_n) = \theta_n(u_n, a(\mathbf{p}), b(\mathbf{p})) + \phi_n(\mathbf{p}) \tag{2.18}$$

where $a(\mathbf{p}), b(\mathbf{p})$ and $\phi(\mathbf{p})$ are homogenous of degree 1 in prices, $\theta_n(\ )$ is homogenous in $a(\mathbf{p})$ and $b(\mathbf{p})$ and, $\sum_n \phi_n(\mathbf{p}) = 0$. If we sum this expenditure function over N, the representative consumers function is

$$e(\mathbf{p}, u_o) = \frac{1}{N} \sum_n c_n(\mathbf{p}, u_n) \tag{2.19}$$

and the average budget share (given by partial of log expenditure with respect to the partial of log price of good j) is

$$\bar{w}_j = \frac{\partial \ln \theta}{\partial \ln a} \frac{\partial \ln a}{\partial \ln p_j} + \frac{\partial \ln \theta}{\partial \ln b} \frac{\partial \ln b}{\partial \ln p_j}. \tag{2.20}$$

Since $\theta_n(\ )$ is homogenous degree 1 in $a(\mathbf{p})$ and $(\mathbf{b})$, we have

$$\frac{\partial \ln \theta}{\partial \ln b} = 1 - \frac{\partial \ln \theta}{\partial \ln a} \tag{2.21}$$

and hence $\bar{w}_j$ can be written as

$$\bar{w}_j = (1 - \lambda) \frac{\partial \ln a}{\partial \ln p_j} + \lambda \frac{\partial \ln b}{\partial \ln p_j} \qquad \text{where } \lambda = \frac{\partial \ln \theta}{\partial \ln b} = \lambda(\mathbf{p}, y_0). \tag{2.22}$$

The term $\lambda(\mathbf{p}, y_0)$ signifies that $\partial \ln \theta / \partial \ln b$ is a function of u_o and $\mathbf{p}$ because θ is a function of u_0 and $\mathbf{p}$ and we can substitute u_0 using the indirect utility function $\psi(\mathbf{p}, y_0)$. The expenditure function (2.18) is called the Generalized Gorman Polar Form and the overall approach is referred to as Generalized Linearity (GL). This approach goes beyond the usual formulation of $y_0 = \bar{y}$ and allows one to incorporate features of expenditure distribution into the demand functions rather than just its mean value. Deaton and Muellbauer consider a special case, in which the representative consumers expenditure level (income) y_0 is assumed to depend on the distribution of individual expenditures (incomes) but not on prices, which leads to particularly useful class of demand equations. If the representative expenditure is independent of prices, then the individual expenditure functions take the form

$$e_n(\mathbf{p}, u_n) = k_n \cdot [a(\mathbf{p})^\alpha (1 - u_n) + b(\mathbf{p})^\alpha u_n]^{1/\alpha}, \tag{2.23}$$

where k_n is a constant that varies over individuals, and α is a constant that is the same for everyone. With the expenditure function above, the budget share equations are said to have the **price independent generalized linear form** (PIGL). For the representative consumer, the expenditure function takes the form (k_n is normalized to one for the reference person)

$$e(\mathbf{p}, u_0) = [a(\mathrm{p})^\alpha (1 - u_0) + b(\mathbf{p})^\alpha u_0]^{1/\alpha}, \tag{2.24}$$

and as $\alpha \to 0$, the representative expenditure function becomes (called PIGLOG)

$$\ln(e(\mathbf{p}, u_0)) = (1 - u_0)\ln(a(\mathbf{p})) + u_0 \ln(b(\mathbf{p})). \tag{2.25}$$

Differentiating log of expenditure function with respect to the log of price of good j gives the nonlinear Engel curves as

$$w_j = \begin{cases} \gamma_j + \eta_j (y/k)^{-\alpha} & \text{PIGL} \\ \gamma_j^* + \eta_j^* \ln(y/k) & \text{PIGLOG} \end{cases} \tag{2.26}$$

where γ's and η's are functions of prices only, k varies over individuals (or households) and can be used to capture demographic effects. Thus, by using a demand/Engle curve that is derived from a PIGL/PIGLOG expenditure function, we are assured that the market demand functions have the same desirable properties as the individual demand functions.

Almost Ideal Demand System (AIDS). The PIGL/PIGLOG family generates exact nonlinear aggregation over individuals or households with nonlinear Engel curves. The merits of representation of market demand as if they were the outcome of decisions by a rational representative consumer has made for extensive application of this class of models. A specific application comes from a second-order Taylor series expansion of equation (2.25) so that the first and second derivatives of the expenditure function with respect to prices and utility (i.e., $\partial e/\partial p_j$,$\partial e/\partial u$, $\partial^2 e/\partial p_j p_i$,$\partial^2 e/\partial u p_j$ and $\partial^2 e/\partial u^2$) can be set equal to those of any arbitrary expenditure function at any point. This is called a flexible functional form. To this end, Deaton and Muellbauer suggest functional forms for $a(\mathbf{p})$ and $b(\mathbf{p})$ in (2.25) which result in a flexible system they call the 'almost ideal demand system', where

$$\ln a(p) = \alpha_0 + \sum_j \alpha_j \ln p_j + \frac{1}{2}\sum_j \sum_k \gamma^*_{jk} \ln p_j \ln p_k$$
$$\ln b(p) = \ln a(p) + \beta_0 \prod_j p_j^{\beta_j} \tag{2.27}$$

so that the AIDS expenditure function is given by

$$\ln e(\mathbf{p}, u) = \alpha_0 + \sum_j \alpha_j \ln p_j + \frac{1}{2}\sum_j \sum_k \gamma^*_{jk} \ln p_j \ln p_k + u\beta_0 \prod_j p_j^{\beta_j} \tag{2.28}$$

with parameters α_j, β_j, and γ^*_{jk}. The expenditure function will be linearly homogenous in $\mathbf{p}$ as long as $\sum_j \alpha_j = 1, \sum_j \gamma^*_{kj} = \sum_k \gamma^*_{kj} = \sum_j \beta_j = 0$. The budget shares of good j can be derived in the usual way (partial of log expenditure with respect to partial of price of good j) which gives the AIDS demand functions in budget share form as

$$w_j = \alpha_j + \sum_k \gamma_{jk} \ln p_k + \beta_j \ln(y/P)$$

where P is a price index defined by (2.29)

$$lnP = \alpha_0 + \sum_k \alpha_k \ln p_k + \frac{1}{2}\sum_i \sum_k \gamma_{ki} \ln p_k \ln p_i$$

and where $\gamma_{jk} = \frac{1}{2}(\gamma^*_{jk} + \gamma^*_{kj})$. The restrictions on the parameter of the cost function impose restriction on the parameters of the AIDS demand system (2.29) given by

$$\sum_{j=1}^{J} \alpha_j = 1 \qquad \sum_{j=1}^{J} \gamma_{jk} = 0 \qquad \sum_{j=1}^{J} \beta_j = 0$$
$$\sum_k \gamma_{jk} = 0 \qquad \gamma_{jk} = \gamma_{kj} \tag{2.30}$$

Provided the restrictions above hold (or are imposed), (2.29) represents a system of demand functions which add up to total expenditure ($\sum w_j = 1$), are homogeneous of degree zero in prices and total expenditure taken together, and satisfy Slutsky symmetry and give nonlinear Engle curves.

2.2. Separability and Multi-Stage Budgeting

The key idea is to solve the dimensionality problem by dividing products into smaller subgroups and allowing flexible substitution between them. In order to make this consistent with theory, we need two related, but different assumptions on consumer preferences regarding separability and multi-stage budgeting. Briefly, separability refers to the case when a consumer's preferences for products of one group are independent of product specific consumption of products from other groups. Multi-stage budgeting refers to when a consumer (or household) can allocate their total expenditure on different goods in sequential stages, represented as a utility tree, where in the first stage, the total current expenditure is allocated to broad groups of products (food, housing, entertainment) followed by allocation of expenditures within each broad group (e.g., meats, vegetables, etc. within the food group). We discuss further each of these below.

Separability. Preferences for products of one group are independent of product specific consumption of products from other groups. Thus,

$$u(q_1, \ldots, q_j) = f[v_1(\mathbf{q_{(1)}}), \ldots, v_k(\mathbf{q_{(k)}}), \ldots v_K(\mathbf{q_{(K)}})], \tag{2.31}$$

where $(q_1, \ldots, q_j) = (\mathbf{q_{(1)}}, \mathbf{q_{(2)}}, \ldots, \mathbf{q_{(k)}})$ i.e., the set $\{\mathbf{q_{(j)}}\}$ is a partition of $(q_1, \ldots, q_j)$ and there are $K < J$ partitions and $f(\cdot)$ is an increasing function of sub-utility functions $v_1, \ldots, v_k$ defined over the partitions. The groups could be broad categories such as food, shelter, etc. or within a class of related products it could be sub groups such as type of food (meat, vegetables, etc.). Note that this does not remove the dimensionality problem but does lessen it. For example, for a linear demand system, the total number of parameters reduces from $J^2 + J$ (additional J parameters are for income) to $J^2/K + K^2$ number of parameters (for $J = 20$ products and $K = 10$ subgroups, we go from a total of 420 parameters to 140 parameters).

While the number of parameters are reduced, separability does impose restrictions on substitution patterns between products in different groups. For the utility function given in (2.31), the implied subgroup demand functions – conditional demand functions – for all products j in group G are of the form

$$q_j = g(y_g, \mathbf{p_g}), \tag{2.32}$$

where $y_g = \sum_{i \in G} p_i q_i$ is the total expenditure on products in group G and $\mathbf{p_g}$ is the vector of prices of these products. Let $s_{ij} = \partial q_i^h / \partial p_j$ be the terms of the Slutsky matrix (i.e., partials of the hicksian demand function with respect to prices), then for any two product $i \in G$ and $j \in H$ where $H \neq G$,

$$\begin{aligned} s_{ij} &= \mu_{GH} \frac{\partial q_i}{\partial y} \frac{\partial q_j}{\partial y} \\ &= \lambda_{GH} \frac{\partial q_i}{\partial y_g} \frac{\partial q_j}{\partial y_h} \\ \text{where } \lambda_{GH} &= \mu_{GH} \frac{\partial y_g}{\partial y} \frac{\partial y_h}{\partial y}. \end{aligned} \tag{2.33}$$

The quantity μ_{GH} summarizes the interrelation between groups and shows that the assumption of weak separability across two groups of goods, plus the assumption that expenditures on both groups of goods rises as total expenditures increases, implies that for any two goods in two different groups, good i in group G and good j in group H, all possible pairs of such goods are either substitutes or complements. For example, if one group is fruit and another group is dairy products, and fruits and dairy products are normal goods, then either all fruit and all dairy products are complements or all fruit and all dairy products are substitutes. The proportionality factor λ_{GH} is the compensated derivative of expenditure on group G with respect to a proportional change in all prices in group H (i.e., $\lambda_{GH} = \sum_{j \in H} p_j \frac{\partial y_g}{\partial p_j}\Big|_{u=const}$), and is thus the intergroup substitution term when each group is defined as a Hick's aggregate with fixed relative prices within the group. If there are K total groups, then we can write a $K \times K$ matrix from the $\lambda's$ that is interpretable as the Slutsky substitution matrix of the group aggregates. Thus, weak separability results in a two-tier structure of substitution matrices: there are K completely general intragroup Slutsky matrices with no restrictions on substitutions within each group, but between groups substitution is limited by (2.33).

The above form of separability is called the weak form. When the marginal rate of substitution between any two goods belonging to the same group is independent of the consumption of goods within the other groups, it is consider as weak separability of preferences. By contrast, if the marginal rate of substitution between any two goods belonging to two different groups is independent of the consumption of any good in any third group, this separability is called strong separability or block additivity.[4] The strong form is when

$$u(q_1, \ldots, q_j) = f[v_1(\mathbf{q_{(1)}}) + \ldots + v_k(\mathbf{q_{(k)}}) + \ldots + v_K(\mathbf{q_{(K)}})], \tag{2.34}$$

and $f'(\cdot) > 0$. In turn, the equivalent form of (2.33) is given by

$$s_{ij} = \mu \frac{\partial q_i}{\partial x} \frac{\partial q_j}{\partial x}, \tag{2.35}$$

where note that μ is independent of groups to which i and j belong.

Multi-stage Budgeting. Consumers can allocate total expenditures in stages, starting with the top level group and then to any subgroups or sub-subgroups within them. At each stage, information appropriate for that stage only is required, i.e., the allocation decision is a function of only that group's total expenditure and price indexes for the subgroups and not of prices or price indexes of products in the other groups. Thus, if the first stage consists of broad categories (food, housing, entertainment) then the consumer decides how much of the budget to allocate to each of these categories depending on three price indexes and not individual prices of types of food items etc. Then, within the food category, the consumer decides how much to spend on different food items (or subgroups) based on the total amount allocated for food and prices of individual food items (or price indexes if there are further subgroups with the food group). Similarly, allocations are done within other groups (housing, entertainment) and the process repeats at a third level if there are subgroups (for instance, within foods group, may have subgroups of meat, vegetables, etc., and then within any of these subgroups there are individual items).

[4]Note that some authors refer to this form as just 'additive' separability (without the use of the word block), but technically that is the case when there is only one good in each group.

Thus the consumer can allocate the expenditures to the subgroups in sequential stages. However, all these sequential allocations must equal those that would occur if the consumers utility maximization problem was done in one complete information step. Because expenditure allocation to any good within a group can be written as a function only of the total group expenditure and the prices of goods within that group, the demand for any good belonging to the group must also be expressed as a function only of total expenditures on the group and the prices of goods within the group.

Weak separability and multi-stage budgeting are closely related concepts but are not the same nor does one imply the other. However, weak separability is necessary and sufficient for the last stage of multi-stage budgeting: separable preferences do not imply multi-stage budgeting but the last stage does imply weak separability. Thus, if a subset of products appear only in a separable sub-utility, then the quantities demanded for these products can be written as a function of expenditure on the group and the prices of individual products within the group. As an example, say there are only three food items (meat, vegetables and drinks) and these food items in the overall utility function appear only as part of separable utility function

$$u(q_1, \ldots, q_m, q_v, q_d, \ldots, q_j) = f[v_1(\mathbf{q_{(1)}}), \ldots, v_F(q_m, q_v, q_d), \ldots, v_K(\mathbf{q_{(K)}})]. \tag{2.36}$$

Then, if the consumer maximizes the above utility function $u(\ )$ subject to the budget constraint, it must be that $v_1, v_2, v_f, \ldots, v_k$ are each maximized subject to the amounts spent on groups $1, 2, \ldots$ (if this were not so, it would mean that $v_1, v_2, v_F, \ldots, v_k$ could be increased without violating the budget constraint and since $f[\ ]$ is an increasing function, it means that utility was not maximized to start with). Hence the expenditures on individual components of the food group (meat, vegetables and drinks) must be the outcome of maximizing $v_F(q_m, q_v, q_d)$ subject to $p_m q_m + p_v q_v + p_d q_d = y_F$ which gives the demand for any specific food item as

$$q_j = g(y_F, p_m, p_v, p_d) \qquad j \in \{m, v, d\}, \tag{2.37}$$

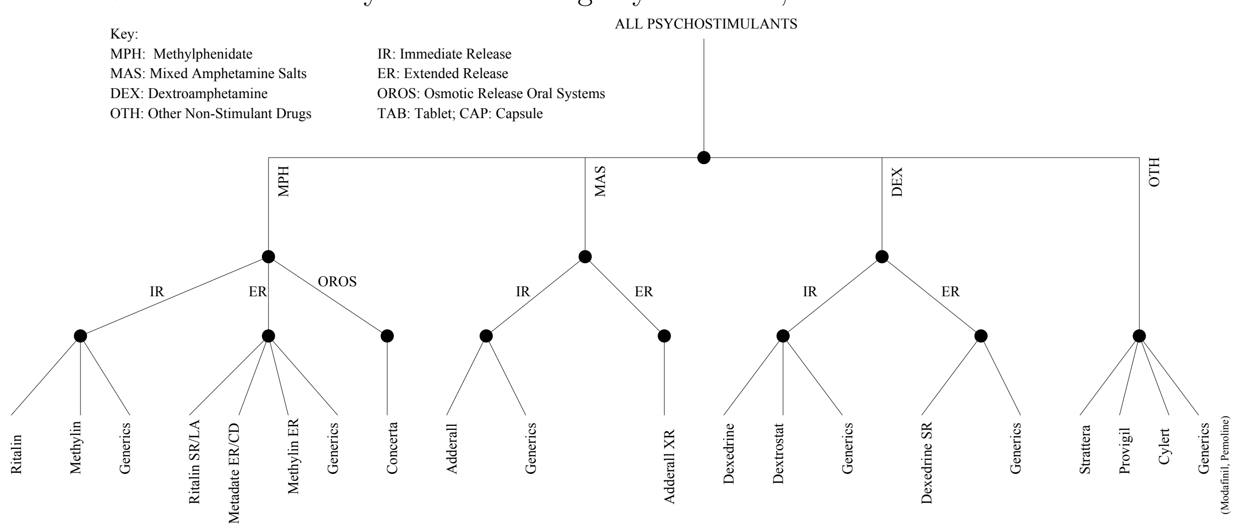
where y_F is the total expenditure on the food items. Conversely, one can also show that existence of these subgroup demand functions implies weak separability.

While weak separability is necessary and sufficient for the last stage of multi-stage budgeting, and one can proceed with group specific demand functions as above, the allocation of total budget to different groups at higher stages requires further restrictions on preferences, or on stronger notions of separability and on composite commodity theorem. To be able to do upper level allocation, there must be an aggregate quantity and price index for each group which can be calculated without knowing the choices within the group. A useful set of requirements is that (1) the overall utility is separably additive in the sub-utilities, and that (2) the indirect utility functions for each group are of the generalized Gorman polar form.

3. Estimation Details with Multistage Budgeting

Hausman, Leonard, and Zona (1994), Hausman (1996), Ellison et al (19xx) use multistage budgeting to construct a multilevel demand system for differentiated products. These applications involve a three-stage system where the top level corresponds to overall demand for the product (beer, pharmaceutical drugs or RTE cereals in the three papers above). The middle level consists of demand for different market segments. For instance, in the demand for beer, the middle segment consists of four groups of beer – premium beer, light beer, imported beer and non-premium beer, while in the RTE cereal paper, the middle segments are family, kids and adults cereals. The bottom level segment involves a flexible brand demand system corresponding to the competition between the different brands within each segment. Similarly, Bokhari and Fournier (2013) use a four stage system where the top level consists of aggregate demand for drugs used in the treatment of ADHD. The second level segments by the types of molecules used in different drugs (four different groups of molecules). The third level further segments the market by the form of the drug, i.e., if it is 4hr, 8hr or a 12hr effect drug. Finally, at the bottom level, different brands and generics are considered within each molecule-form segment of the market. See figure below.

Figure 1. Taxonomy of ADHD drugs by Molecule, Form and Brand Names



Note: Generics refer to several manufacturers for each molecule and form given in the column. There are no generic versions of Concerta and Adderall XR during the study period.

3.1. Specifications

For each of these stages a flexible parametric functional form is assumed. The choice of functional form is driven by the need for flexibility, but also requires that the conditions for multistage budgeting are met.

Bottom Level. A typical application has the AIDS model at the lowest level. The demand for product i in segment fm, which consists of I_{fm} number of products, in area a at period t is given by

Level 1 (Bottom):

$$s_{iat_{fm}} = \alpha_{i_{fm}} + \beta_{i_{fm}} ln(\frac{R_{fmat}}{P_{fmat}}) + \sum_{j=1}^{I_{fm}} \gamma_{ij_{fm}} \ln P_{jat_{fm}} + \mathbf{x}_{iat_{fm}} \boldsymbol{\lambda}_{i_{fm}} + \varphi_{iat_{fm}} \tag{3.1}$$

where $s_{iat_{fm}}$ is the revenue share of product i (i.e., share within segment fm in market at) and is a function of I_{fm} number of prices (log prices of other products in the segment, $\ln P_{jat_{fm}}$), total expenditure and a price index for the segment fm (R_{fmat} and P_{fmat}) and other exogenous variables ($\mathbf{x}_{iat_{fm}}$) which may be varying by product, market or segment and may include terms like demographic variables, time trends, area fixed effects or any observable product characteristics if they vary by markets (if we include area and time dummies, then $\alpha_{i_{fm}}$ can be alternatively written as $\alpha_{iat_{fm}}$). Note that there are as many equations as the number of products in segment fm, and one has to estimate a system of such equations for each segment, either jointly (all equations from all segments together) or on a segment-by-segment basis if for some reason data does not allow joint estimation of all segments. As discussed earlier, this system defines a flexible functional form that can allow for a wide variety of substitution patterns within the segment. It has two additional advantages over other flexible demand systems (like the Rotterdam system or the Translog model): (1) it aggregates well over individuals; and (2) it is easy to impose (or test) theoretical restrictions, like adding-up, homogeneity of degree zero and symmetry. To impose the restrictions, we require (for each segment)

$$\begin{aligned} &\sum_{i=1}^{I_{fm}} \alpha_{i_{fm}} = 1 \qquad \sum_{i=1}^{I_{fm}} \gamma_{ik_{fm}} = 0 \qquad \sum_{i=1}^{I_{fm}} \beta_{i_{fm}} = 0 \\ &\sum_{k} \gamma_{ik_{fm}} = 0 \qquad \gamma_{ik_{fm}} = \gamma_{ki_{fm}} \end{aligned} \tag{3.2}$$

where the last share equation per segment is not estimated as the shares must add up to one (recall that the revenue shares are shares relative to total spending in this segment and not total spending on all drugs).

Deaton and Muellbaur's exact price index P_{fmat} is given by

$$lnP_{fmat} = \alpha_{0_{fm}} + \sum_{i}^{I_{fm}} \alpha_{i_{fm}} \ln P_{iat_{fm}} + \frac{1}{2} \sum_{i}^{I_{fm}} \sum_{k}^{I_{fm}} \gamma_{ki_{fm}} \ln P_{kat_{fm}} \ln P_{iat_{fm}} \tag{3.3}$$

and since it involves the parameters that need to be estimated, AIDS estimation requires non-linear estimation methods. In practice however, Deaton and Muelllbaur suggest using the **Stone price index**

$$\ln P_{fmat} = \sum_{i}^{I_{fm}} s_{iat_{fm}} \ln P_{iat_{fm}} \tag{3.4}$$

which makes estimation much simpler – setting aside the issue of endogeniety due to the correlation of prices and the error term in equation (3.1), which we discuss later – the system of equations can be estimated via OLS or SUR. In general, researchers have often found that results do not seem to be too sensitive regardless of whether the exact price index is used or the Stone index is used (this does not mean that one should not check within their own application).

Note however that by using the Stone price index, even if the original prices were not endogenous, we are introducing an artificial endogeneity, as equation (3.1) now involves shares on both the left and right hand side of the equation. To overcome this difficulty, Hausman and colleagues suggest using an area specific average value of $s_{iat_{fm}}$ in the Stone index construction in (3.4), where the average is over the multiple years. To be clear, in (3.4), they replace $s_{iat_{fm}}$ with $\bar{s}_{ia_{fm}}$ so that the value is different for each city but the same for all periods. In these applications, the choice of a city-specific average over time is dictated, in part, due to the fact that the data consists of a few cities but over many periods (monthly or quarterly observations for a few years). Bokhari and Fournier (2013), instead use $\bar{s}_{it_{fm}}$, i.e., period specific average, so that each time period has different value, but is the same for all cities. This choice is dictated by two reasons. First, they have the opposite situation in terms of observations with annual sales data for four years but disaggregated by many geographic areas (counties in US) and second, because different drugs are introduced in the market in different years, and when they are introduced in a give year, they reach all the geographic areas simultaneously. Taking the average of shares over years for a given geographic area would would require including zero shares in the average for the periods when the drug was not available in the market. Bokhari and Fournier also verify that the results are not too different regardless of if one uses $\bar{s}_{ia_{fm}}$ or $\bar{s}_{it_{fm}}$ if the data is restricted to periods when all drugs are on the market.

Middle level(s). At the next level up (the middle level, or level 2), demand captures the allocation between segments and can again be modeled using the AIDS specification, in which case the demand specified by equation (3.1) is used with both expenditure shares and prices aggregated to a segment level. The prices are aggregated using either equations (3.3) or (3.4). However, if the latter is used, Bokhari and Fournier suggest using $s_{iat_{fm}}$ for the purpose of creating a price index for the upper level rather than $\bar{s}_{at_{fm}}$ or $\bar{s}_{it_{fm}}$. An alternative is the log-log equation used by Hausman, Leonard, and Zona (1994) and Hausman (1996) and is given by

Level 2 (Middle):

$$\ln(Q_{[fm]at}) = A_{[fm]} + B_{[fm]} \ln(R_{at}) + \sum_{n=1}^{FM} \Gamma_{[fm]n} \ln P_{nat} + \mathbf{x}_{[fm]at} \boldsymbol{\lambda}_{[fm]} + \xi_{[fm]at} \tag{3.5}$$

where (suppressing subscripts at for areas and periods) $q_{[fm]}$ is the aggregate quantity of the $[fm]$ bottom level segment, i.e., total quantity of RTE cereals for the family, kids or the adults segments in market at (city and quarter) and $P_{[fm]}$ is the price of each of these $[fm]$ segments, written as $\ln P_n$ in the equation above, where n is an indexing number for the lower level $[fm]$ segment. Thus, these segment level prices are the price indexes from the lower level equations and are computed using equations (3.3) or (3.4) as discussed earlier (with the caveat that we use actual lower level shares rather than their average values over area or periods to pass on to the higher level). The variable R_{at} is the total expenditure by market on all related products. For instance, it is the sum of total sales of RTE cereals over the the three segments, kids, family and adults. Finally, $\mathbf{x}_{[fm]at}$ are the exogenous variables that are segment specific characteristics – if they are different for each market – or just demographic variables by markets. Again, area and period specific fixed effects (or time trends instead of the latter) can be alternatively written as $A_{[fm]at}$. Note that the number of equations to be estimated is equal to the number of lower level segments.

Since the lower level of the demand system is AIDS, which satisfies the generalized Gorman polar form, the preferences of the second level should be additively separable (i.e., overall utility from ready-to-eat cereal or all ADHD drugs should be additively separable in the sub-utilities from the various subsegments), in order to be consistent with exact two-stage budgeting. Neither the second level AIDS, nor the log-log system satisfy this requirement (Deaton and Muellbauer also discuss approximate – instead of exact – two-stage budgeting, and show that if one uses the Rotterdam model, approximate two-stage budgeting implies that higher stages also have Rotterdam functional form). Also, in order for exact multistage budgeting to hold to the next level of aggregation, discussed later, these preferences should be of generalized Gorman polar form.

Before moving on to the discussion of the top level equation, note that Bokhari and Fournier (2013) have two middle level segments which differentiate drugs by forms within molecules (level 2) and by molecules among all ADHD drugs (level 3). At level 2, they again use an AIDS specification and at level 3 they use a log-log specification like the Hausman level 2 equation. Accordingly, their two middle level equations are

Level 2 (Middle):

$$u_{fat_m} = a_{f_m} + b_{f_m} \ln(\frac{R_{mat}}{P_{mat}}) + \sum_{h=1}^{F_m} g_{fh_m} \ln P_{hat_m} + \mathbf{x}_{fat_m}\boldsymbol{\lambda}_{f_m} + \mu_{fat_m} \tag{3.6}$$

Level 3 (Middle):

$$\ln(Q_{mat}) = A_m + B_m \ln(R_{at}) + \sum_{n=1}^{M} \Gamma_{mn} \ln P_{nat} + \mathbf{x}_{mat}\boldsymbol{\lambda}_m + \xi_{mat}.$$

In the level 2 equations above (suppressing subscripts at for exposition), u_{f_m} and P_{h_m} are the revenue share and price of form f within molecule m where, under the Stone price index version, the latter is given by equation (3.4) – $ln(P_{f_m}) = \sum_{j=1}^{I_{fm}} s_{i_{fm}} \ln(P_{j_{fm}})$ – and is the share weighted sum of the log prices of products within the form f. To be clear, this is the price index used in bottom (level 1) equations. However, the terms $\frac{R_m}{P_m}$ are the total expenditures from all forms within molecule m, and a price index for molecule m where the later is computed (using Stone index form) as

$$\ln(P_m) = \sum_{h=1}^{F_m} u_{f_m} \ln(P_{h_m}). \tag{3.7}$$

For level 2, one needs to estimate as many equations as there are forms per molecule (F_m), and repeat the process for each molecule. For instance, if there are four molecules, and each admits up to three forms, then a total of four sets of system equations, with each set consisting of three equations need to be estimated. Again, depending on the data, the estimations can be joint for all segments, or segment by segment, and restrictions can be imposed within each segment much like the lower levels.

Level 3 is an aggregation from level 2, so that $\ln q_m$ is the aggregate quantity for segment m and is the sum of quantities over all forms within this molecule. Similarly, $\ln P_n$ is the price of molecule n used earlier in level 2 and is given by (3.7). Total number of equations to be estimated equals the number of upper level segments, e.g., total number of molecules and the rest is the same as discussed earlier in the context of middle level equation (3.5).

Top level. Finally, the top level is the demand for the entire set of subsegments (RTE cereal, beer, ADHD drugs etc.) and is typically specified as

$$\begin{aligned}&\text{Level 4 (Top):}\\&\ln Q_{at} = A + B\ln(Y_{at}) + G\ln P_{at} + \mathbf{x}_{at}\boldsymbol{\lambda} + \zeta_{at}\end{aligned} \tag{3.8}$$

where q_{at} is the total quantity, Y_{at} is the real income, $\mathbf{x}_{at}$ are the demand shifters and P_{at} is the overall price index for these products, given by share weighted sum of (log) prices at the previous level and given by (again suppressing subscripts at),

$$\ln(P) = \sum_{m=1}^{M} v_m \ln(P_m) \tag{3.9}$$

where v_m is the revenue share and P_m is the price index for molecule m computed earlier in (3.7). Note that this form does satisfy additive separability, which is required for exact two-stage budgeting.

Conditional and Unconditional Elasticities. The multi-budgeting process allows estimation of the conditional demand functions (conditional on expenditures on the segment) at the lower levels and the cross-price elasticities are limited to within the segment. Nonetheless, from these conditional demand estimates, and estimates of the upper level equations, it is possible to derive the unconditional cross-price elasticities across the full range of products in different segments, albeit subject to the cross segment restrictions noted earlier in section 2.2. For the four level multi-stage process above, the conditional and unconditional elasticities are given below.

Conditional on segment expenditure R_{fm} (in market at), price elasticity of a product is

$$\begin{aligned}\frac{\partial lnQ_{i_{fm}}}{\partial lnP_{k_{f'm'}}} &= \frac{1}{s_{i_{fm}}}\Big\{\Big(-\beta_{i_{fm}}\bar{s}_{k_{f'm'}} + \gamma_{ij_{f'm'}}\Big)\cdot 1[f'=f, m'=m]\Big\} \\ &\quad - 1[i=k, f'=f, m'=m],\end{aligned} \tag{3.10}$$

where $1[\cdot]$ is the indicator function. Thus, elasticities conditional on R_{f_m} are zero across products in different f-m segments. Note that the subscript at has been suppressed in the equation above but is present on all quantities, shares, prices etc. and $\bar{s}_{k_{f'm'}}$ is either $\bar{s}_{kt_{f'm'}}$ or $\bar{s}_{ka_{f'm'}}$ depending on whichever one was used in the Stone price index in level 1 share equations. Similarly, elasticity at level 2 with respect to the *price index* for the segment and conditional on segment revenue R_m in market at (where the market subscripts have been suppressed), has a similar formula as for the bottom level (since both are in AIDS form) and is given by

$$\frac{\partial lnQ_{f_m}}{\partial lnP_{f'_{m'}}} = \frac{1}{u_{f_m}}\Big\{\Big(-b_{f_m}\bar{u}_{f'_{m'}} + g_{fh_{m'}}\Big)\cdot 1[m'=m]\Big\} - 1[f'=f, m'=m], \tag{3.11}$$

and the conditional cross price elasticity of forms in different level 3 segments (i.e., for forms in different molecules) is zero. Price elasticities at level 3 (for example, at the molecule level), are just the Γ_{mn} parameters in level 3 equation, and similarly, elasticity with respect to price for the aggregate product is the value of the parameter G in top level equation.

Given the parameters of the conditional demand equations and shares of products, the unconditional elasticity (for the four level system) can be computed as

$$\begin{aligned}\frac{\partial lnQ_{i_{fm}}}{\partial lnP_{k_{f'_{m'}}}} &= \Big(1+\frac{\beta_{i_{fm}}}{s_{i_{fm}}}\Big)\bar{s}_{k_{f'_{m'}}}\Big[\frac{g_{ff'_{m'}}}{u_{fm}} + \bar{u}_{f'_{m'}}\Big]\cdot 1[m=m'] \\ &\quad + \Big(1+\frac{\beta_{i_{fm}}}{s_{i_{fm}}}\Big)\bar{s}_{k_{f'_{m'}}}\Big[\frac{b_{fm}\bar{u}_{f'_{m'}}}{u_{fm}} + \bar{u}_{f'_{m'}}\Big]\Gamma_{mm'} \\ &\quad + \frac{1}{s_{i_{fm}}}\Big\{\gamma_{ik_{f'_{m'}}} - \beta_{i_{fm}}\bar{s}_{k_{f'_{m'}}}\Big\}\cdot 1[f'=f, m'=m] \\ &\quad - 1[i=k, f'=f, m'=m].\end{aligned} \tag{3.12}$$

3.2. Instruments

In an earlier section we discussed how endogeneity can arise in the context of a competitive single product demand-supply model, where due to the simultaneity, the price and the error term in the demand equation are correlated (see equation (1.15)). This basic idea extends to a variety of differentiated products pricing models. Let the demand for the i^{th} product be given by $q_i = D_i(\mathbf{p}, \mathbf{z}_i; \xi_i)$, where ξ_i is the error term and consists of unobserved product characteristics, and $\mathbf{z}_i$ is the vector of exogenous demand shifters (say the observed product characteristics). If there are L firms, and the lth firm produces a subset $\mathfrak{L}_l$ of the products, then it maximizes its joint profit over these products as

$$\Pi_l = \sum_{r\in\mathfrak{L}_l}(p_r - c_r)q_r(\mathbf{p}, \mathbf{z}_r, \xi_r), \tag{3.13}$$

where c_r is the constant marginal cost of the r^{th} product. Under Nash-Bertrand price competition, price p_i of any product i produced by firm l satisfies the first order condition

$$q_i(\mathbf{p}, \mathbf{z}_i; \xi_i) + \sum_{r\in\mathfrak{L}_l}(p_r - c_r)\frac{\partial q_r(\mathbf{p}, \mathbf{z}_r; \xi_r)}{\partial p_i} = 0. \tag{3.14}$$

Then the equilibrium price for product i would be a function of its marginal cost and a markup term, and in matrix form (for all equilibrium prices) is given by

$$\mathbf{p} = \mathbf{c} + \mathbf{\Omega^{-1}q}(\mathbf{p}, \mathbf{z}; \boldsymbol{\xi}), \tag{3.15}$$

where $\mathbf{\Omega}$ is defined such that $\Omega_{ri} = -O_{ri}\frac{\partial q_r(\mathbf{p},\mathbf{z}_r;\xi_r)}{\partial p_i}$ and $\boldsymbol{O}$ is 1/0 joint ownership matrix with ones in the leading diagonals and in r, i position if these products are produced by the same firm and zeros everywhere else. As can be seen, the markup term is a function of the same error terms, and hence generally prices will be endogenous, so that OLS/SUR estimation will lead to biased estimates of the demand parameters.

The usual starting place for demand side instruments is to use cost shifters (terms that affect $\mathbf{c}$, such as cost of raw materials) that are uncorrelated with demand shocks. These can work well for homogenous products, but in the case of differentiated products, we would need costs shifters that vary by individual brands, which are often very difficult to obtain. However, if such cost shifters can be found that are uncorrelated with demand side shocks and they vary by individual brands, then they would be good instruments.

In the face of this difficulty, IO literature has often used two types of instruments which have grown in popularity and are described below (this is not an endorsement to blindly apply these instruments as they may or may not be applicable in a given situation).

The first, due to Berry (1994) builds on Bresnahan's (1981) assumption that the location of products in a characteristics space is determined prior to the revelation of consumer's valuation of the unobserved product characteristics. BLP use this assumption to generate a set of instrumental variables. Specifically, they use the observed product characteristics (excluding price and any other endogenous characteristics of the product), the sums of the values of the same characteristics of other products offered by that firm, and the sums of the values of the same characteristics of products offered by other firms. Consider the case when there are two firms, X and Y and each is producing three products A,B,C and D,E,F respectively. Suppose further that each of these products have two observable characters, S (say, package size, which is the number of pills in a box) and T (number of times a pill must be taken during a day for a standard diagnosis). Then for the price of A, which is produced by firm X, there are 6 potential instruments:

- S_{AX} and T_{AX} – the values of S and T of product A
- $S_{BX} + S_{CX}$ and $T_{BX} + T_{CX}$ – the sum of S and T over the firms two other products B and C
- $S_{DY} + S_{EY} + S_{FY}$ and $T_{DY} + T_{EY} + T_{FY}$ – the sum of S and T over the competitors products D,E and F

Similar instruments can be constructed for prices of other products. The main advantage of this approach (if valid) is that it gives instruments that vary by brands. Nonetheless, problems arise if the assumption that the observed characteristics are uncorrelated with observed characteristics is not valid. This could happen, for instance, if the observed characteristics are changing over time, and the change in observed characteristics is for the same unobserved factors that determine price. Another potential issue arises if brand dummies are included in the estimation, since then it must be the case that there is variation in products offered in different markets, else there will be no variation between the instruments in these markets.

A second set of instruments is due to Hausman et al. (1994) and has been used in several papers. Hausman uses the panel nature of data and the assumption that prices in different areas (cities) are correlated via common cost shocks, to use prices from other areas as instruments for prices in a given city. The identifying assumption is that after controlling for brand specific intercepts and demographics, the city specific valuations of a product are independent across cities but may be correlated within a city over time. Given this assumption, the prices of the brand in other cities are valid instruments so 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 the market specific valuation of the product.

To be clear, the general idea is the reduced form price of a product i in two cities, $a = 1$ and $a = 2$ at time period t, will be given by

$$\begin{aligned} \ln p_{i1t} &= \pi_1 \ln c_{it} + \mathbf{x}_{i1t}\boldsymbol{\pi}_2 + v_{i1t} \\ \ln p_{i2t} &= \pi_1 \ln c_{it} + \mathbf{x}_{i2t}\boldsymbol{\pi}_2 + v_{i2t}, \end{aligned} \tag{3.16}$$

where c_{it} is the common cost component of the price in two different cities and $\mathbf{x}_{iat}$ are brand level demand shifters (demographics, time trends) as well city specific brand differentials (intercepts by brands and cities) due to differences in transportation costs or local wages. In general, the error terms v_{iat} will be correlated with φ_{iat} in the equation (3.1), and hence OLS/SUR will give inconsistent estimates. If however, v_{i1t} is uncorrelated with v_{i2t}, then city two's prices will be uncorrelated with the error term φ_{i1t} in the equation (3.1), and hence the instrument will be valid.

Further, since the prices in the two cities are driven by the same underlying common costs c_{it}, they will be correlated to each other and hence relevant.

In addition to common cost shocks, the Hausman instruments also rely on no correlation between v_{i1t} and v_{i2t}. However, this assumption may be invalid if the terms are related due to common demand side shocks across the two cities. For instance, a national campaign will increase the unobserved valuation of product i in both cities, thus violating the independence assumption.

4. Characteristic Space Approaches

The characteristics space approach starts with a consumer choosing a single product from a finite set of goods. The model defines each product as a bundle of attributes (including price, which is a special attribute), and consumers have preferences over these attributes. Consumers can have different relative preferences, which gives rise to the random coefficients models, and they choose the product that maximizes their utility subject to the usual constraints. This leads to different choices by different consumers. Aggregate demand is then derived as the sum over individuals and depends on the entire distribution of consumer preferences.

General Approach. Indirect utility for individual n for product j in market t is given by

$$u_{njt} = U(\mathbf{x}_{jt}, \xi_{jt}, y_{nt} - p_{jt}, \boldsymbol{d}_{nt}, \boldsymbol{\nu}_{nt}, \epsilon_{njt}; \boldsymbol{\theta}_n), \qquad \text{for } j = 0, 1, 2, \ldots, J. \tag{4.1}$$

In the equation above, 0 refers to the 'outside good', chosen when the consumer does not purchase any of the observed products. Price of the outside good is often considered to be exogenous or not known and set to zero. The vector $\mathbf{x}_{jt}$ and random variable ξ_{jt} are the observed and unobserved (to the econometrician, but not to the consumer) product characteristics and do not vary over consumers. The product characteristics, multiplied by the parameters $\boldsymbol{\theta_n}$ determine the level of utility for consumer n. The vectors $\boldsymbol{d}_{nt}$ and $\boldsymbol{\nu}_{nt}$ are vectors of observed and unobserved sources of differences in consumer tastes. They do not enter the utility function directly, but rather enter into the model by changing the value of the parameters of interest for each consumer. For instance, $\boldsymbol{d}_{nt}$ may be a vector of observed demographics (income, family size etc.), that effect the parameters (marginal valuations) of product characteristics by individual, and change the value of $\boldsymbol{\theta}$ for each attribute of the product by individual n. Similarly, for each product attribute (including price) there is an additional randomness to the marginal valuation by individuals and is captured by $\boldsymbol{\nu}_{nt}$. This term is added into the model because there may be unobserved person specific characteristics that affect their marginal valuation for an observed product characteristic. A specific example would be the number of dogs a family owns that affects their marginal valuation of the size of a car that they want to purchase, i.e., unobserved number of dogs owned by a family changes the coefficient on car size which is a variable part of $\mathbf{x}_{jt}$. Note that if $\mathbf{x}_{jt}$ is a $k-1$ vector of observed characteristics, then $\boldsymbol{\nu}_{nt}$ is a vector of length k (the additional dimension is for price). Thus, the coefficients $\boldsymbol{\theta}_n$ depend on $\boldsymbol{d}_{nt}$ and $\boldsymbol{\nu}_{nt}$. Additionally, ϵ_{njt} is a mean-zero stochastic term that enters directly into the utility of product j for consumer n. This term captures the idiosyncratic variation in consumer preferences by individual products, and effects the level of (normalized) utility associated with a specific product j and varies by individuals. Just to be clear, note that for each consumer, $\boldsymbol{\epsilon}_{nt} = (\epsilon_{n0t}, \epsilon_{n1t}, \ldots, \epsilon_{nJt})$ is a vector of error terms with the length equal to the number of products. This term does not affect the value of the parameters $\boldsymbol{\theta}_n$ (that's what $\boldsymbol{d}_{nt}$ and $\boldsymbol{\nu}_{nt}$ are for). Finally, y_{nt} is the consumers income, but is often subsumed into either $\boldsymbol{\nu}$ or in $\boldsymbol{d}$, so that utility is modelled explicitly depending on prices, i.e., $u_{njt} = U(\mathbf{x}_{jt}, \xi_{jt}, p_{jt}, \boldsymbol{d}_{nt}, \boldsymbol{\nu}_{nt}, \epsilon_{njt}; \boldsymbol{\theta}_n)$. Utility of the outside good is denoted as $u_{n0t} = U(\mathbf{x}_{0t}, \xi_{0t}, \boldsymbol{d}_{nt}, \boldsymbol{\nu}_{nt}, \epsilon_{n0t}; \boldsymbol{\theta})$ and is normalized to zero.

Consumers choose products that give them the highest utility. Thus, consumer n will choose product j when $u_{njt} \geq u_{nlt}$ for all $l = 0, 1, \ldots, J$ and $l \neq j$. Since the differences in consumer choices arise only due to differences in the marginal valuations $\boldsymbol{\theta}_n$ (which are themselves functions of $\boldsymbol{d}_{nt}$ and $\boldsymbol{\nu}_{nt}$), and the idiosyncratic terms ϵ_{njt}, a consumer can be described as a tuple $(\boldsymbol{d}, \boldsymbol{\nu}, \boldsymbol{\epsilon})$ which then defines a set of individual attributes that lead to the choice of good j, given by

$$\mathbb{A}_{jt}(\mathbf{x}_t, \mathbf{p}_t; \boldsymbol{\theta}) = \{(\boldsymbol{d}_{nt}, \boldsymbol{\nu}_{nt}, \epsilon_{n0t}, \epsilon_{n1t}, \ldots, \epsilon_{nJt}) \mid u_{njt} > u_{nlt} \qquad \forall\, l = 0, 1, 2 \ldots J,\ l \neq j\}. \tag{4.2}$$

where $\mathbf{p}_t = (p_{0t}, \ldots, p_{Jt})'$ and $\mathbf{x}_t = (\mathbf{x}_{0t}, \ldots, \mathbf{x}_{Jt})'$. The set $\mathbb{A}_{jt}$ defines characteristics of the individuals that choose brand j in market t. If there are no ties (or the probability of ties is zero), then the market share of product j is just the probability weighted sum of individuals in the set $\mathbb{A}_{jt}$. Thus, if $F(\boldsymbol{d}, \boldsymbol{\nu}, \boldsymbol{\epsilon})$ is the population joint distribution function, then the market share of product j is the integral of this distribution over the mass of individuals in the region $\mathbb{A}_{jt}$,

$$s_{jt}(\mathbf{x}, \mathbf{p}; \boldsymbol{\theta}) = \int_{\mathbb{A}_{jt}} dF(\boldsymbol{d}, \boldsymbol{\nu}, \boldsymbol{\epsilon}). \tag{4.3}$$

If the size of the market is M (total number of consumers) then the aggregate demand for the jth product is $M s_{jt}(\mathbf{x}, \mathbf{p}; \boldsymbol{\theta})$.

These will become clearer when we look at the specific cases. We consider two models below.

5. Standard Logit/Homegenous Tastes

Let the indirect utility for consumer n for product j in market t be given by

$$\begin{aligned} &u_{njt} = \alpha_n(y_n - p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta}_n + \xi_{jt} + \epsilon_{njt}, \text{ where} \\ &n = 1, \ldots, N, \qquad j = 0, 1, \ldots, J, \qquad t = 1, 2, \ldots, T, \text{ and} \\ &\boldsymbol{\beta}_n = \boldsymbol{\beta}, \qquad \alpha_n = \alpha, \qquad \text{for all } N. \end{aligned} \tag{5.1}$$

In this model we are assuming that there is no variation in tastes across consumers and the terms $\boldsymbol{d}_{nt}$ and $\boldsymbol{\nu}_{nt}$ do not enter this model (but later on will make $\boldsymbol{\beta}_n$ and α_n functions of $\boldsymbol{d}_n$ and $\boldsymbol{\nu}_n$ mentioned earlier). The vector $\mathbf{x}_{jt}$ is a $k-1$ dimensional vector of observable characteristics (which may be varying by markets) and ξ_{jt} is a *scalar* that summarizes the unobservable (to the econometrician) product characteristics, and neither of these terms varies over consumers. Thus, if there are multiple unobserved characteristics, then we are assuming that they can be collapsed into a single index whose value does not vary over consumers.

Equation (5.1) is an indirect utility function which can be derived from a quasilinear utility function of the form $U(\mathbf{q}) = q_0 + u_1(q_1) + \ldots + u_J(q_J)$ which is maximized subject to the usual budget constraint ($\sum_{j=0}^{J} p_j q_j \leq y_n$) and that $\sum_{j=0}^{J} q_j = 1$. The outside option (product 0) is normalized by assuming that the price and other characteristics are zero for this option so that

$$u_{n0t} = \alpha y_n + \epsilon_{n0t}. \tag{5.2}$$

The utility function in (5.1) can be written more compactly as just

$$u_{njt} = \alpha y_n + \delta_{jt} + \epsilon_{njt}, \tag{5.3}$$

where $\delta_{jt} \equiv \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt}$ is the mean utility for product j in market t. Since income is common to all options, and consumers only differ in the terms ϵ, the set of individuals choosing product j is given by

$$\mathbb{A}_{jt}(\boldsymbol{\delta}_t(\mathbf{x}_t, \mathbf{p}_t; \alpha, \boldsymbol{\beta})) = \{(\epsilon_{n0t}, \epsilon_{n1t}, \ldots \epsilon_{nJt}) | u_{njt} > u_{nlt} \qquad \forall\ l = 0, 1, 2 \ldots J,\ l \neq j\}. \tag{5.4}$$

where $\boldsymbol{\delta}_t = (\delta_{0t}, \ldots, \delta_{Jt})'$, and $\mathbf{p}_t$ and $\mathbf{x}_t$ are defined as before. For the logit model, we assume that ϵ_{njt} are independently and identically distributed (iid) and follow a Type-1 extreme value distribution, given by

$$f(\epsilon) = \exp(-\epsilon)\exp(-\exp(-\epsilon)) \quad \text{and} \qquad F(\epsilon) = \exp(-\exp(-\epsilon)), \tag{5.5}$$

where $f(\epsilon)$ and $F(\epsilon)$ are the PDF and CDF of the random variable ϵ. In this case, the market share of product j (and the probability that individual n chooses product j) is

$$s_{jt}(\boldsymbol{\delta}_t) = \int_{\mathbb{A}_{jt}} dF(\boldsymbol{\epsilon}) = \frac{\exp(\delta_{jt})}{\sum_{j=0}^{J} \exp(\delta_{jt})}. \tag{5.6}$$

Note that income has dropped out of the equation for shares as it was common to all options. Due to the earlier normalization $\delta_{0t} = 0$ (so that $(\exp(\delta_{0t}) = \exp(0) = 1)$, the share equation above can be written as

$$\begin{aligned} s_{jt} &= \frac{\exp(\delta_{jt})}{1 + \sum_{j=1}^{J} \exp(\delta_{jt})} \\ s_{0t} &= \frac{1}{1 + \sum_{j=1}^{J} \exp(\delta_{jt})} = 1 - \sum_{j=1}^{J} s_{jt}. \end{aligned} \tag{5.7}$$

Thus, $s_{jt}/s_{0t} = \exp(\delta_{jt})$, and hence

$$\ln(s_{jt}) - \ln(s_{0t}) = \delta_{jt} \equiv \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt} \tag{5.8}$$

can be estimated using linear regression methods. The dimensionality problem mentioned earlier, i.e., that one needs to estimate parameters on the order of J^2 to compute the full cross elasticity matrix, has been reduced to estimation of just one parameter α, as the own and cross elasticities can be computed using observed shares, prices and the value of α. The closed (logit) form for the shares is due to both, the extreme value distribution, and the iid assumption. The latter, especially the independence part of iid, causes serious limitations on the substitution patterns. Both the estimation details and the limitations of the logit model are discussed below.

5.1. Elasticities and Substitution Patterns

The logit model suffers from the property known as the Independence of Irrelevant Alternatives (IIA). To understand this, it is easiest to go back to the individual level probabilities of choosing a product. When the error term is iid and has extreme value distribution (I am dropping the subscript t for exposition), the probability that individual n chooses product j is given by (see (5.6))

$$\Pr(j) = \frac{\exp(\delta_j)}{\sum_{j=0}^{J} \exp(\delta_j)}. \tag{5.9}$$

The relative probabilities of options j and k are thus

$$\frac{\Pr(j)}{\Pr(k)} = \frac{\exp(\delta_j)}{\exp(\delta_k)} = \exp(\delta_j - \delta_k). \tag{5.10}$$

This ratio does not depend on characteristics of any other alternative other than those of j and k. This implies that the relative odds of choosing j over k are the same no matter what other alternatives are available or what are the attributes of other alternatives (the values of $\delta' s$). Since the ratio is independent of the alternatives other than j and k, it is said to be independent of irrelevant alternatives (and hence the term IIA).

In some situations this property is unrealistic (Chipman, 1960; Debreu 1960). Consider the famous example of blue bus/red bus: A traveler can commute to work either by car (c) or by blue bus (bb). Suppose further that it turns out (for simplicity) that $Pr(bb) = Pr(c) = .5$. Now suppose that a new type of bus is introduced that is identical in all other respects to the existing blue bus (fare, route, smell, time it takes to get to work, etc.,) except that it is red in color (rb). One would expect that the new probabilities of travel model would be such that $Pr(bb) = Pr(rb) = .25$ and $Pr(c) = .5$. Yet the logit model would predict that the substitution from the two old modes of travel (blue bus or car) to the new mode of travel (red bus) are such that they would depend on the ratio of old probabilities, and since they were equal, the new probabilities for each of the new modes would be $Pr(bb) = Pr(rb) = Pr(c) = 1/3$. This aggregated example exemplifies the implications of IIA in a logit model.

This same issue embodies itself in terms of cross-elasticities of logit probabilities with respect to any of the characteristics of the choices under consideration (see pp. 49-52 in Train). To see its impact in terms of aggregate demand (based on logit model), compute the own- and cross-price elasticities of shares in equation (5.6) or equation (5.8) with respect to prices. It is straight forward to see that

$$\eta_{jkt} = \frac{\partial s_{jt}}{\partial p_{kt}} \frac{p_{kt}}{s_{jt}} = \begin{cases} -\alpha p_{jt}(1 - s_{jt}) & \text{if } j = k, \\ \alpha p_{kt} s_{kt} & \text{otherwise.} \end{cases} \tag{5.11}$$

Note that the cross price elasticity between product j and k depends only on the prices and shares of product k. In turn, it implies that the cross price elasticity of any two products - say product j which is Coca Cola and product l which is Orange Cola, with respect to the price of product k which is Pepsi Cola will be the same. Put another way, the logit model predicts that if the price of Pepsi Cola increases by 1%, then ceteris paribus, the market shares of Coca Cola and Orange Cola will increase by the same proportion regardless of the fact that Coca Colas and Pepsi Cola are more like each other (blue bus/red bus) compared to Orange Cola (car). Clearly this is unrealistic for many markets.

An additional issue with the logit-based elasticities is that of own price elasticity (or elasticity with respect to any other characteristic). In most markets, the shares of any given product are likely to be small, for instance if there are many differentiated products, or if the size of the potential market is large. In this case, the own elasticity will be roughly proportional to the price of the product ($\eta_{jjt} \approx -\alpha p_{jt}$ because $(1 - s_{jt}) \approx 1$). This means that if price increases, sensitivity to prices also increases – but people who buy more expensive products may infact be less price sensitive compared to those who buy less expensive products. Similarly, if as the price increases, so does elasticity, it implies that the markups for cheaper priced products will be larger than those with higher priced products (price-costs margin inversely related to own elasticities). Again, this may not be true in some industries, for instance, in pharmaceuticals. Do we believe that markups are higher for cheaper priced generics compared to the blockbuster patented and higher price drugs?

While the logit is clearly unrealistic in terms of implied substitution patterns, the main question prior to shunning such a model is to ask, if in fact, you even need a more realistic substitution pattern to answer the research question at hand. For example, what if you don't care about the cross-price elasticities as long as you can get decent estimates of own-price elasticities, and the industry is such that the criticism about own elasticities above does not apply? In this case, the ease of logit estimation may be quite appealing, and even if it is not, since it is cheap to estimate it, it can still serve as a useful starting point for more elaborate estimations, which we consider next.

5.2. **Estimation Details**

Outside Good. If we have aggregate sales data (quantities and prices), along with product characteristics, equation (5.8) can be estimated by defining the dependent variable y_{jt} as $y_{jt} = \ln(s_{jt}) - \ln(s_{0t})$. However, to operationalize it, we need to estimate the share of the outside good. This is done by first defining the (potential) size of the market. Researchers have used different definitions for it. Nevo (2001) defines the potential size of the market as one bowl of cereal per day per person, BLP define it as total number of households, while Bresnahan et al (1997) define it as the total number of office-based employees. In the example of ADHD drugs considered earlier, one could define it as a 12-hr day-long coverage of a standard dose of ADHD drug – 3 × 30mg strength of Ritalin IR (a 30mg pill covers about 4hrs of a day) which can be multiplied by a base line candidate population, say 10% of all school aged children (current ADHD prevalence rates of whom only 69% are given any ADHD drugs), and a smaller proportion of the older population. Once the potential size of the market M_t is defined, then based on the observed values of $q_{1t}, \ldots, q_{Jt}$, one can define the shares of the 'inside' goods $s_{1t}, \ldots s_{Jt}$ relative to the market size as

$$s_{jt} = q_{jt}/M_t \qquad j = 1, \ldots, J \text{ for all } t = 1, \ldots, T. \tag{5.12}$$

Then, share of the outside good per market is just

$$s_{0t} = 1 - \sum_{j=1}^{J} s_{jt} \qquad \forall t. \tag{5.13}$$

With these definitions in place, estimation of equation (5.8), which I rewrite below,

$$\ln(s_{jt}) - \ln(s_{0t}) = \delta_{jt} \equiv \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt}, \tag{5.8}$$

is in principle, straight forward and can be estimated via OLS (ignoring endogeneity issues for the moment) with data from even just <u>one</u> market. To see this more clearly, it is helpful to write the equation in matrix form. Let $\mathbf{y}_t' = (y_{1t}, y_{2t}, \ldots, y_{Jt})$ be a row vector (for market t) given by $\mathbf{y}_t' = ([\ln s_{1t} - \ln s_{0t}], [\ln s_{2t} - \ln s_{0t}], \ldots, [\ln s_{Jt} - \ln s_{0t}])$ so that $\mathbf{y}_t$ is a column vector of length J. Similarly, let $\mathbf{p}_t' = (p_{1t}, \ldots, p_{Jt})$ and $\boldsymbol{\xi}_t' = (\xi_{1t}, \ldots, \xi_{Jt})$ be row vectors with J entries for the t^{th} market. We already defined $\mathbf{x}_{jt}$ as a row vector of observable characteristics of product j in market t, i.e., $\mathbf{x}_{jt} = (x_{1jt}, x_{2jt}, \ldots, x_{Kjt})$, thus let $\mathbf{X}_t' = (\mathbf{x}_{1t}', \mathbf{x}_{2t}', \ldots, \mathbf{x}_{jt}', \ldots, \mathbf{x}_{Jt}')$ so that $\mathbf{X}_t$ is a $J \times K$ matrix, such that each row is itself a k dimensional vector of observable product characteristics.[5]

[5] In an earlier section I stated that $\mathbf{x}_{jt}$ is a $k-1$ dimensional vector. However, here I am using it as k-dimensional vector. This is because I do not want to write subscripts in the following equations that go up to $K-1$, as the notation becomes very cumbersome, and the equations too wide to fit easily on a page. Thus, for the next few equations, pretend as if there were K observable characteristics (excluding price) and not $K-1$. Alternatively, where ever K appears in the subscript in the next two equations, replace it with $K-1$.

Then (5.8) can be written in 'long' form as

$$\mathbf{y}_t = (\ln \mathbf{s}_{jt} - \ln s_{0t}) \qquad = \alpha(-\mathbf{p}_t) + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\xi}_t \equiv \boldsymbol{\delta}_t$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_J \end{bmatrix}_t = \begin{bmatrix} \ln s_1 - \ln s_0 \\ \ln s_2 - \ln s_0 \\ \vdots \\ \ln s_J - \ln s_0 \end{bmatrix}_t = \alpha \begin{bmatrix} -p_1 \\ -p_2 \\ \vdots \\ -p_J \end{bmatrix}_t + \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ \vdots & & & \\ x_{J1} & x_{J2} & \dots & x_{JK} \end{bmatrix}_t \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_J \end{bmatrix}_t \tag{5.14}$$

where subscript t outside the matrices is to highlight the matrix estimation within each market. As long as the number of product characteristics is only a handful, say five, compared to the number of total products, say 50, then the number of parameters to be estimated is seven ($\boldsymbol{\beta}, \alpha$ and variance of the error term in (5.8) to compute the variance-covariance matrix). This is not to suggest that you should use data from just a single market, only that estimation is possible with just one market and the form above shows more clearly how to organize the data in long form for logit estimation. When we have data from multiple markets, long form data from each market can be vertically 'stacked'. Thus, equation above can be written in matrix notation as

$$\mathbf{y} = \alpha(-\mathbf{p}) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \equiv \boldsymbol{\delta}$$

$$\begin{bmatrix} \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{J1} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{Jt} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} y_{1T} \\ y_{2T} \\ \vdots \\ y_{JT} \end{pmatrix} \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} \ln s_{11} - \ln s_{01} \\ \ln s_{21} - \ln s_{01} \\ \vdots \\ \ln s_{J1} - \ln s_{01} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \ln s_{1t} - \ln s_{0t} \\ \ln s_{2t} - \ln s_{0t} \\ \vdots \\ \ln s_{Jt} - \ln s_{01} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \ln s_{1T} - \ln s_{0T} \\ \ln s_{2T} - \ln s_{0T} \\ \vdots \\ \ln s_{JT} - \ln s_{0T} \end{pmatrix} \end{bmatrix} = \alpha \begin{bmatrix} \begin{pmatrix} -p_{11} \\ -p_{21} \\ \vdots \\ -p_{J1} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} -p_{1t} \\ -p_{2t} \\ \vdots \\ -p_{Jt} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} -p_{1T} \\ -p_{2T} \\ \vdots \\ -p_{JT} \end{pmatrix} \end{bmatrix} + \begin{bmatrix} \begin{pmatrix} x_{111}\, x_{121} \dots x_{1K1} \\ x_{211}\, x_{221} \dots x_{2K1} \\ \vdots \\ x_{J11}\, x_{J21} \dots x_{JK1} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{11t}\, x_{12t} \dots x_{1Kt} \\ x_{21t}\, x_{22t} \dots x_{2Kt} \\ \vdots \\ x_{J1t}\, x_{J2t} \dots x_{JKt} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{11T}\, x_{12T} \dots x_{1KT} \\ x_{21T}\, x_{22T} \dots x_{2KT} \\ \vdots \\ x_{J1T}\, x_{J2T} \dots x_{JKT} \end{pmatrix} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \begin{pmatrix} \xi_{11} \\ \xi_{21} \\ \vdots \\ \xi_{J1} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \xi_{1t} \\ \xi_{2t} \\ \vdots \\ \xi_{Jt} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \xi_{1t} \\ \xi_{2t} \\ \vdots \\ \xi_{Jt} \end{pmatrix} \end{bmatrix} \tag{5.15}$$

where $\mathbf{y}, \mathbf{p}$ and $\boldsymbol{\xi}$ are $JT \times 1$ vectors, $\mathbf{X}$ is a $JT \times K$ matrix, α is a scaler and $\boldsymbol{\beta}$ is $K \times 1$ vector.

Instruments and Dummy Variables. The error term ξ_{jt} in equation (5.8) consists of the the unobserved (to the econometrician) product characteristics reduced to a single index value. As discussed in section 3.2 before, this term is likely to be correlated with the prices so that $\text{cov}(p_{jt}, \xi_{jt}) \neq 0$. As before, one needs to find instruments that are correlated with price but not with any of the unobserved product characteristics – and without repeating the earlier discussion – one can try the Hausman instruments (prices from other markets) and/or the BLP instruments (sum of observed product characteristics of own-firm other products, and of sum of product characteristics from competitors) if they are suitable for the problem at hand.

Regardless of the instruments used, a first approach to consistent estimation would be to estimate a fixed effects model with dummies for products (and markets). This of course requires that data be available from multiple markets. Thus, with data available from multiple markets, one can estimate via OLS

$$\ln(s_{jt}) - \ln(s_{0t}) = \delta_{jt} = \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_j + \xi_t + \Delta\xi_{jt} \tag{5.16}$$

where ξ_j is the brand fixed effect and ξ_t is the market fixed effect. In this case the identifying assumption for OLS estimation is

$$\mathrm{E}(\Delta\xi_{jt} p_{jt}|\mathbf{x}_{jt}) = 0. \tag{5.17}$$

A brand specific dummy variable captures all the observed characteristics of the product that do not vary across markets, as well as the product specific mean of the unobserved characteristics, i.e., $\mathbf{x}_j\boldsymbol{\beta}$, where, note the missing market subscript of t from the vector $\mathbf{x}$. Thus, the correlation between prices and brand specific mean of unobserved quality is fully accounted for and does not require an instrument. Once brand specific dummy variables are included in the regression, the error term now is just the market specific deviation from the mean of the unobserved characteristics, and may still require the use of instruments if the condition in equation (5.17) is not true.

Similarly, if the mean unobserved quality – where the mean is now across all brands – is different by markets, then it too is fully accounted for by the market dummies. For instance, if the subscript t for the markets is in the context of time periods, then this could be because the unobserved quality for all products is improving over time (think computer quality over time). If the subscript t is in the cross-sectional setting, then this may or may not make much sense, since by adding such dummies to the equation, the researcher is effectively arguing that the unobserved quality components of all brands in, Hooker, OK, are higher than those in Boring, OR. This maybe true if the products under study require some additional local input for providing the product (radio channels with local DJs and ads), or if shipping from long distance affects the quality of all products (fresh food), but not if they are centrally produced (RTE cereals) and shipping does not impact quality.

Note that the use of brand dummies increases the number of parameters to be estimated by J (rather than by J^2), and may not be too serious an issue if the number of markets is large. A potentially more serious difficulty is that the coefficients $\boldsymbol{\beta}$ cannot be identified if observed characteristics do not vary by markets. Nevo (2001) points out that in fact they can be recovered using minimum distance procedure by regressing the estimated brand dummy variables on the observed characteristics. For instance, let $\mathbf{b}_t$ be the $J \times 1$ vector of brand dummies and let $\mathbf{X}_t$ be the $J \times K$ matrix of observed product characteristics and $\boldsymbol{\xi}_t$ be the $J \times 1$ vector of unobserved product qualities, neither of which vary by markets. Let also $\hat{\boldsymbol{b}}$ be the estimated values of coefficients ($J \times 1$) of the brand dummies and

$\hat{\boldsymbol{V}}_{\boldsymbol{b}}^{-1}$ their estimated $J \times J$ variance covariance matrix, both of which are available from initially estimating equation (5.16), either via OLS, or IV if prices were treated as endogenous. Then, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ in equation

$$\mathbf{b}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\xi}_t, \tag{5.18}$$

can be recovered via GLS estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}_t'\hat{\mathbf{V}}_{\mathbf{b}}^{-1}\mathbf{X}_t)^{-1}\mathbf{X}_t'\hat{\mathbf{V}}_{\mathbf{b}}^{-1}\hat{\mathbf{b}}_t, \text{ and } \widehat{\boldsymbol{\xi}}_t = \hat{\mathbf{b}}_t - \mathbf{X}_t\widehat{\boldsymbol{\beta}} \tag{5.19}$$

where the latter is just the calculated value of the residual term from the regression above.

5.3. Nested Logit

One problem we saw with the simple logit model is related to the cross price elasticities, i.e., consumers substitute towards other brands in proportion to the market shares. Intuitively, if the price of a product goes up, we would expect them to substitute towards other products that are more similar to the original product. This IIA problem arose from the iid structure of the error term in the logit model. Particulary, while consumers have different rankings of the products, these differences arise only due to the iid shocks to the error term ϵ_{njt}.

A solution to this problem is to make the random shocks to the utility correlated across products by generating correlations through the error term, so as to effectively get rid of the independence component. An example is the nested logit model in which products are grouped and ϵ_{njt} is decomposed into an iid shock plus a group specific component which results in correlation between products in the same group. Thus, the basic idea is to relax the IIA by grouping products (similar to the grouping idea in multilevel budgeting/AIDS we saw earlier), but within each group we have a standard logit model, and products in different groups have less in common and are not good substitutes.

Formally, the utility for consumer n for product j in group g is given by

$$u_{njt} = \delta_{jt} + \zeta_{ngt}(\sigma) + (1-\sigma)\epsilon_{njt}, \tag{5.20}$$

where, as before, $\delta_{jt} = \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt}$ is the mean utility for product j common to all consumers, ϵ_{njt} is still the person specific iid random shock with extreme value distribution, but ζ_{ngt} is the person specific shock that is common to all products in group g. The distribution of the group specific random variable ζ_{ngt} depends on the parameter σ so that $\zeta_{ngt}(\sigma) + (1-\sigma)\epsilon_{njt}$ is extreme value. If σ approaches zero, the model is reduced to that of the simple logit case discussed earlier while if it approached one, only the nests matter. As discussed in Berry (1994), the utility function (5.20) leads to a closed form logit equation for shares (much like the earlier one we saw) with one additional term, and is given by

$$\ln(s_{jt}) - \ln(s_{0t}) = \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \sigma \ln(s_{jt}/s_{gt}) + \xi_{jt}. \tag{5.21}$$

In the equation above, the additional term $\ln(s_{jt}/s_{gt})$ is the share of product j in group g. The methods/issues vis-a-via definition of the outside good, use of product dummies, and endogeneity of prices are all the same as those discussed earlier and hence won't be repeated here. However, one difference from the previous case is that even if prices are exogenous, the term $\ln(s_{jt}/s_{gt})$ is clearly endogenous and hence one needs to find an additional variable to instrument for this term.

5.4. Review of Generalized Methods of Moments (GMM)

Before turning to the more general model, the random coefficients logit, it is worth reviewing GMM estimation as we will use it in a later section.

Suppose we want to estimate a simple linear model

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + u_t, \tag{5.22}$$

where $\mathbf{x}_t$ is a $1 \times K$ vector (including the constant or the intercept term), $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters and u_t is the usual error term. Suppose further that conditional on the values of the regressors, the error term is mean zero so that $\mathrm{E}[u_t|\mathbf{x}_t] = 0$. This single conditional moment restriction leads to K conditional moment conditions $\mathrm{E}[\mathbf{x}_t' u_t] = \mathbf{0}$ due to the law of iterated expectations. Thus, if the error has conditional zero mean, we get K equations of the form

$$\mathrm{E}[\mathbf{x}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta})] = \mathbf{0}. \tag{5.23}$$

The method of moments (MM) estimator is the solution to the corresponding sample moment conditions

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta}) = \mathbf{0}, \tag{5.24}$$

which gives the MM estimator of $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}_{MM} = (\sum_t \mathbf{x}_t'\mathbf{x}_t)^{-1} \sum_t \mathbf{x}_t' y_t = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{5.25}$$

which is just the OLS estimator.

Consider next the case where we still want to estimate the equation (5.22), but where the error terms are correlated with the variables $\mathbf{x}_t$, but that we have an additional set of exogenous variables $\mathbf{z}_t$ that are correlated with $\mathbf{x}_t$ but not with the error terms so that $\mathrm{E}[u_t|\mathbf{z}_t] = 0$. Then, $\mathrm{E}[(y_t - \mathbf{x}_t\boldsymbol{\beta})|\mathbf{z}_t] = 0$, and as before, we can multiply $\mathbf{z}_t$ with the residual terms to get K unconditional population moment conditions

$$\mathrm{E}[\mathbf{z}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta})] = \mathbf{0}. \tag{5.26}$$

In this case the MM estimator solves the sample moment conditions given by

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta}) = \mathbf{0}, \tag{5.27}$$

and if $\dim(\mathbf{z}) = K$, then this yields the MM estimator which is just the IV estimator

$$\widehat{\boldsymbol{\beta}}_{MM} = (\sum_t \mathbf{z}_t'\mathbf{x}_t)^{-1}\sum_t \mathbf{z}_t' y_t = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{5.28}$$

If however, $\dim(\mathbf{z}) > K$, (more potential instruments than the original number of regressors) then there is no unique solution as there are more moment conditions than the number or parameters to be estimated. In this case, the GMM estimator kicks in and chooses $\widehat{\boldsymbol{\beta}}$ in such as way so as make the vector $T^{-1}\sum_{t=1}^{T}\mathbf{z}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta})$ as small as possible using quadratic loss. Thus, the GMM method finds $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$ which minimizes the function

$$Q(\boldsymbol{\beta}) = \left[\frac{1}{T}\sum_t \mathbf{z}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta})\right]' \boldsymbol{\Phi} \left[\frac{1}{T}\sum_t \mathbf{z}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta})\right] \tag{5.29}$$

where $\boldsymbol{\Phi}$ is a $\dim(\mathbf{z}) \times \dim(\mathbf{z})$ weighting matrix.

In matrix notation define $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ (where $\mathbf{y}$ and $\mathbf{u}$ are $T \times 1$, $\mathbf{X}$ is $T \times K$ and $\boldsymbol{\beta}$ is $K \times 1$ as before), and let $\mathbf{Z}$ be $T \times R$ matrix, then $\sum_{t=1}^{T}\mathbf{z}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta}) = \mathbf{Z}'\mathbf{u}$ and (5.29) becomes

$$Q(\boldsymbol{\beta}) = \left[\frac{1}{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{Z}\right] \boldsymbol{\Phi} \left[\frac{1}{T}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \tag{5.30}$$

where $\boldsymbol{\Phi}$ is a $R \times R$ full rank symmetric weighting matrix. To solve for $\boldsymbol{\beta}$ we can compute and solve for the first order conditions, $\partial Q(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$. In the forgoing case of linear IV, the first order conditions are

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2\left[\frac{1}{T}\mathbf{X}'\mathbf{Z}\right] \boldsymbol{\Phi} \left[\frac{1}{T}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] = \mathbf{0}. \tag{5.31}$$

These lead to the GMM linear IV estimator and its variance

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{GMM}} &= (\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\mathbf{y} \\ \mathrm{V}(\widehat{\boldsymbol{\beta}})_{\text{GMM}} &= T(\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\mathbf{Z})^{-1}\Big(\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\widehat{\mathbf{S}}\boldsymbol{\Phi}\mathbf{Z}'\mathbf{X}\Big)(\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\mathbf{X})^{-1},\end{aligned} \tag{5.32}$$

where $\widehat{\mathbf{S}}$ is a consistent estimate of

$$\mathbf{S} = \text{plim}\frac{1}{T}\sum_i\sum_j\left[\mathbf{z}_i' u_i u_j \mathbf{z}_j\right]. \tag{5.33}$$

Different choices of the weighting matrix $\mathbf{\Phi}$ lead to different estimators. If the model is just identified ($R = K$) and the matrix $\mathbf{X}'\mathbf{Z}$ is invertible, then the choice of the weighting matrix $\mathbf{\Phi}$ does not matter as the GMM estimator is just the IV estimator:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{GMM}} &= \left(\mathbf{X}'\mathbf{Z}\mathbf{\Phi}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{\Phi}\mathbf{Z}'\mathbf{y} \\ &= \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{\Phi}^{-1}\left(\mathbf{X}'\mathbf{Z}\right)^{-1}(\mathbf{X}'\mathbf{Z})\mathbf{\Phi}\mathbf{Z}'\mathbf{y} \\ &= \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y} = \widehat{\boldsymbol{\beta}}_{\text{IV}}.\end{aligned} \tag{5.34}$$

If $R > K$, and the errors are homoscedastic, then $\mathbf{\Phi} = (T^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$ and $\widehat{\mathbf{S}}^{-1} = [s^2T^{-1}\mathbf{Z}'\mathbf{Z}]$ leads to the usual 2SLS estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{GMM}} &= \left(\mathbf{X}'\mathbf{P}_{\mathbf{z}}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{P}_{\mathbf{z}}\mathbf{y}\right) = \widehat{\boldsymbol{\beta}}_{\text{2SLS}} \\ \text{V}(\widehat{\boldsymbol{\beta}}_{\text{GMM}}) &= s^2\left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1} \\ \text{where } \mathbf{P}_{\mathbf{z}} &= \mathbf{Z}\left(\mathbf{Z}\mathbf{Z}'\right)^{-1}\mathbf{Z}' \text{ and } s^2 = (T-K)^{-1}\sum_t \hat{u}_t^2,\end{aligned} \tag{5.35}$$

and if the errors are heteroscedastic, then instead we can use

$$\begin{aligned}\text{V}(\widehat{\boldsymbol{\beta}}_{\text{GMM}}) &= T\left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\widehat{\mathbf{S}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)\left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1} \\ \text{and } \widehat{\mathbf{S}} &= T^{-1}\sum_t \hat{u}_t^2\mathbf{z}_t\mathbf{z}'_t.\end{aligned} \tag{5.36}$$

So far GMM is the same as 2SLS or IV. The **optimal** weighting matrix (optimal in the sense of efficiecy/smallest variance) is one which is proportional to the inverse of $\mathbf{S}$. The optimal GMM two-step estimator (for the linear IV case) is when $\mathbf{\Phi} = \widehat{\mathbf{S}}^{-1}$

$$\widehat{\boldsymbol{\beta}}_{\text{OGMM}} = \left(\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{y}, \tag{5.37}$$

where one has to first figure out a consistent value of $\widehat{\mathbf{S}}^{-1}$. One way to proceed is to use the 2SLS as the first-step to estimate $\widehat{\boldsymbol{\beta}}$ to compute the residuals as in the heteroscedastic case above, and then proceed to construct the $\widehat{\mathbf{S}}^{-1}$ and then use it in (5.37) to compute the estimator. The variance of the optimal GMM estimator is then given by

$$\text{V}(\widehat{\boldsymbol{\beta}}_{\text{OGMM}}) = T\left(\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}. \tag{5.38}$$

Often though, in the computation of $\text{V}(\widehat{\boldsymbol{\beta}}_{\text{OGMM}})$, one use an alterative version of $\widehat{\mathbf{S}}$, say $\tilde{\mathbf{S}}$, which is computed as given in (5.36) but with the small difference that the residuals used in that equation are not from those from the end of 2SLS estimation, but rather the residuals that result post computing the optimal GMM estimator in (5.37).

This approach extends easily to the general case with other moment conditions as well. Let $\boldsymbol{\theta}$ be a $q \times 1$ vector of parameters and $\mathbf{h}(\mathbf{w}, \boldsymbol{\theta})$ be an $r \times 1$ vector function such that at the true value of the parameter $\boldsymbol{\theta}_0$, there are r moment conditions ($r > q$) give by

$$\text{E}\left[\mathbf{h}(\mathbf{w}_t, \boldsymbol{\theta}_0)\right] = \mathbf{0}, \tag{5.39}$$

and that the expectations are not zero if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. The vector $\mathbf{w}_t$ includes all observable variables, including $\mathbf{y}_t, \mathbf{x}_t$ and, $\mathbf{z}_t$. Then the GMM objective function (equivalent of (5.29)) is

$$Q(\boldsymbol{\beta}) = \left[\frac{1}{T}\sum_t \mathbf{h}(\mathbf{w}_t, \boldsymbol{\theta})\right]' \boldsymbol{\Phi} \left[\frac{1}{T}\sum_t \mathbf{h}(\mathbf{w}_t, \boldsymbol{\theta})\right], \tag{5.40}$$

and the corresponding first order conditions are

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left[\frac{1}{T}\sum_t^T \frac{\partial \mathbf{h}_t(\widehat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}}\right] \boldsymbol{\Phi} \left[\frac{1}{T}\sum_t^T \mathbf{h}_t(\widehat{\boldsymbol{\theta}})\right] = \mathbf{0}, \quad \text{where } \mathbf{h}_t(\boldsymbol{\theta}) = \mathbf{h}(\mathbf{w}_t\boldsymbol{\theta}). \tag{5.41}$$

If $\mathbf{h}_t(\boldsymbol{\theta}) = \mathbf{z}_t'(y_t - \mathbf{x}_t\boldsymbol{\beta}) = \mathbf{z}_t'u_t$ then $\partial\mathbf{h}/\partial\boldsymbol{\beta}' = -\mathbf{z}_t'\mathbf{x}_t$ and the earlier results of linear IV follows.

Similarly, GMM also extends to non-linear models, where the error term u_t may or may not be additively separable. For instance, $u_t = y_t - g(\mathbf{x}_t; \boldsymbol{\theta})$ where $g(\cdot)$ is some nonlinear function but the error term is additively separable, or non-separable so that $u_t = g(y_t, \mathbf{x}_t; \boldsymbol{\theta})$. Either way, if $\mathrm{E}(u_t|\mathbf{x}_t) \neq 0$ but we have instruments available so that $\mathrm{E}(u_t|\mathbf{z}_t) = 0$, then the moment conditions that follow are $\mathrm{E}(\mathbf{z}_t'u_t) = \mathbf{0}$ and the GMM estimator minimizes the objective function

$$Q(\boldsymbol{\beta}) = \left[\frac{1}{T}\mathbf{u}'\mathbf{Z}\right] \boldsymbol{\Phi} \left[\frac{1}{T}\mathbf{Z}'\mathbf{u}\right]. \tag{5.42}$$

Unlike the linear case, the first order conditions do not give closed forms for the estimators.

5.5. Back to Logits

We saw earlier that standard logit can be estimated as a linear equation when the dependent variable is defined as $y_{jt} \equiv \ln s_{jt} - \ln s_{0t}$ and the equation is given as $y_{jt} = \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt}$. When the price is correlated with the unobserved heterogeneity term ξ_{jt}, so that $\mathrm{E}(p, \xi) \neq 0$ and we have a set of instruments such that $\mathrm{E}(Z\xi) = 0$, then we can use the GMM/IV methods described in the earlier section to estimate the parameters of the equation. For instance, equipped with a set of instruments, this could be done in a straight forward way using equations (5.37) and (5.38). However, the linear equation arose out of Berry's (1994) inversion trick. It is worth while to go back to the basic logit probability model and work it out (again) as we will need to use a more detailed version of this basic trick when we do not have closed forms given above.

To start with, note that in each market t we can observe the actual shares of each product (i.e., that is part of our data). Let the actual observed shares be given by $\mathbf{s}$ so that $\mathbf{s}_t = (s_{0t}, s_{1t}, \ldots, s_{Jt})$ where, as before, $s_{0t} = 1 - \sum_{j=1}^{J} s_{jt}$ is the share of the outside good. Next, if have a probability/market share model – see equation (5.7) – and if we know the values of the model parameters $\alpha, \boldsymbol{\beta}$ (henceforth let $\boldsymbol{\theta}_1 \equiv \begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}'$), then we can always predict market shares based on equation (5.7). Let the model predicted market shares be given by $\tilde{\mathbf{s}}$ so that $\tilde{\mathbf{s}}_t = (\tilde{s}_{0t}, \tilde{s}_{1t}, \ldots, \tilde{s}_{Jt})$.

Then we can use the guiding principle of finding values of $\boldsymbol{\theta}_1$ so that $\mathbf{s}_t$ is as close as possible to $\tilde{\mathbf{s}}_t$. Thus, we may be tempted to use methods of non-linear least squares (NLS) to find $\boldsymbol{\theta}_1$ that minimizes the distance between the observed and predicted market shares,

$$\min_{\boldsymbol{\theta}_1} \sum_{j=1}^{J} \left[s_{jt} - \tilde{s}_{jt}(\alpha, \boldsymbol{\beta}, \xi_{1t}, \xi_{2t}, \ldots, \xi_{Jt})\right]^2 \tag{5.43}$$

in every market t, i.e., minimize the sum of squared residuals between the observed and the predicted market shares. The problem with this method in general is that the econometric error terms $\boldsymbol{\xi}_t$ – unobserved product qualities – enter the predicted market share and are not additively separable. Hence, non-linear least squares methods will not give consistent estimates *even if* prices were not endogenous.

Berry (1994) suggest using the GMM framework with a transformation that makes errors additively separable. Assume that we have a set of M instruments given by matrix $\mathbf{Z}$ with dimensions $JT \times M$ (the jt^{th} row is given by $\mathbf{z}_{jt} = (z_{jt}^{(1)}, z_{jt}^{(2)}, \ldots, z_{jt}^{(M)})$) which are uncorrelated with error terms in the utility model ξ_{jt}. Then the M moment conditions are given by $\mathrm{E}(\mathbf{z}'_{jt}\xi_{jt}) = \mathbf{0}$. Note that even if there was only one market, the the moment conditions hold, i.e., in that case the expectation is over the J products within the single market. The key insight comes from the fact that the error terms enter the mean utility linearly ($\delta_{jt} = \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt}$), and that they only enter the mean utility and hence one can separate out the ξ_{jt} terms to compute the moment conditions above. The sample analog for the m^{th} moment in market t is given by

$$\frac{1}{J}\sum_{j} z_{jt}^{(m)}\xi_{jt} = \frac{1}{J}\sum_{j} z_{jt}^{(m)}(\delta_{jt} - \mathbf{x}_{jt}\boldsymbol{\beta} + \alpha p_{jt}). \tag{5.44}$$

Thus we want to estimate the parameters $\alpha, \boldsymbol{\beta}$ that minimize the sample moment conditions (or rather their weighted sum of squares). But since we cannot observe δ_{jt} we cannot proceed as is. To this end, Berry (1994) suggests a two step approach. In the first step, we obtain an estimate of δ_{jt}, – call it $\widehat{\delta}_{jt}$ – and insert it into the moment conditions above, and in the next step, we search for values of $\alpha, \boldsymbol{\beta}$ that minimize the weighted sum of squares of these moment conditions.

(1) Figure out the values of δ_{jt}

  (a) If we normalize $\delta_{0t} = 0$ and equate the observed shares to model predicted shares, then we have J non-linear equations per market – see logit share equation (5.7) – in J unknowns

$$\begin{aligned} s_{1t} &= \tilde{s}_{1t}(\delta_{1t}, \ldots, \delta_{Jt}) \\ s_{2t} &= \tilde{s}_{2t}(\delta_{1t}, \ldots, \delta_{Jt}) \\ &\vdots \\ s_{Jt} &= \tilde{s}_{jt}(\delta_{1t}, \ldots, \delta_{Jt}). \end{aligned} \tag{5.45}$$

(b) If we can invert this system, we can solve for $\delta_{1t}, \delta_{2t}, \ldots, \delta_{jt}$ as a function of observed shares $s_{1t}, s_{2t}, \ldots, s_{jt}$.

(c) Thus, we now have $\widehat{\delta}_{jt} \equiv \tilde{s}_{jt}^{-1}(s_{1t}, s_{2t}, \ldots, s_{Jt})$, J numbers per market which we can use to carry out step 2 (in the simple logit case, $\widehat{\delta}_{jt} = \ln(s_{jt}) - \ln(s_{0t})$)

(2) With the estimated values of δ_{jt}, use GMM to estimate parameters (in this case, α and $\boldsymbol{\beta}$) so as to minimize (5.44).

(a) Recall that δ_j is the mean utility of product j defined linearly as $\delta_{jt} = \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt}$ for all j,

$$\begin{aligned} \delta_{1t} &= \alpha(-p_{1t}) + \mathbf{x}_{1t}\boldsymbol{\beta} + \xi_{1t} \\ \delta_{2t} &= \alpha(-p_{2t}) + \mathbf{x}_{2t}\boldsymbol{\beta} + \xi_{2t} \\ &\vdots \\ \delta_{Jt} &= \alpha(-p_{Jt}) + \mathbf{x}_{Jt}\boldsymbol{\beta} + \xi_{Jt}. \end{aligned} \tag{5.46}$$

(b) We can now use the estimated values of $\widehat{\delta}_j$ to calculate the sample moments

$$\frac{1}{J}\sum_j z_{jt}^{(m)}\xi_{jt} = \frac{1}{J}\sum_j z_{jt}^{(m)}(\widehat{\delta}_{jt} - \mathbf{x}_{jt}\boldsymbol{\beta} + \alpha p_{jt}) \tag{5.47}$$

and minimize these to calculate the values of $\alpha, \boldsymbol{\beta}$.

In the case of the simple logit, we saw earlier that δ_{jt} has a closed analytical form and is simply equal to $\ln s_{jt} - \ln s_{0t}$ – log of the observed market shares minus the log of share of the outside good – and hence we can straight away proceed to step 2, i.e., set the moment conditions to $\frac{1}{J}\sum_j z_{jt}^{(m)}(\ln s_{jt} - \ln s_{0t} - \mathbf{x}_{jt}\boldsymbol{\beta} + \alpha p_{jt})$ to find $\boldsymbol{\theta}_1 = \begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}'$. This comes back to just estimating the equation $\ln s_{jt} - \ln s_{0t} = \mathbf{x}_{jt}\boldsymbol{\beta} - \alpha p_{jt} + \xi_{jt}$ via IV/2SLS/GMM techniques discussed earlier. An analytical form is also available for some of the other cases (nested logit, for instance). More generally however, as in the random coefficients model discussed next, this will not be so. I outline the points of departure below.

In step (1a) above, we equated observed market shares to model predicted market shares. In the case of logits, the model predicted market shares take the closed form (5.7) given by $\tilde{s}_{jt} = \exp(\delta_{jt})/\left[1 + \sum_{j=1}^{J}\exp(\delta_{jt})\right]$. In other cases, there will be no closed form available to compute the model predicted market shares and we will need to resort to numerical simulation methods to estimate the model predicted shares. Moreover, these will be functions of additional parameters (call them $\boldsymbol{\theta}_2$). Thus, equations (5.45) will be of the form

$$s_{jt} = \tilde{s}_{jt}(\delta_{1t}, \ldots, \delta_{Jt}, \boldsymbol{\theta}_2) \tag{5.48}$$

Next, in steps (1b/1c), we 'inverted' these equations to solve for $\widehat{\delta}_{jt}$. In the case of logit, analytical solution was available since $\delta_{jt} = \ln s_{jt} - \ln s_{0t}$. More generally, these equations are nonlinear and need to be solved numerically. Berry/BLP suggest a contraction mapping (and prove that it converges) for $\boldsymbol{\delta}_t$ given by

$$\boldsymbol{\delta}_t^{h+1} = \boldsymbol{\delta}_t^h + \left[\ln(\mathbf{s}_t) - \ln(\tilde{\mathbf{s}}_t(\boldsymbol{\delta}_t^h; \boldsymbol{\theta}_2))\right], \tag{5.49}$$

where $\mathbf{s}_t(\cdot)$ is the observed market share, $\tilde{\mathbf{s}}_t(\cdot)$ is the model predicted market share at mean utility $\boldsymbol{\delta}_t^h$ at iteration h and $||\boldsymbol{\delta}_t^{h+1} - \boldsymbol{\delta}_t^h||$ is below some tolerance level.

Thus, in the random coefficients model that follows, we will again use Berry's two step method (as for the simple logit) with the exception that we will have to compute model predicted market shares using simulation methods and that the inversion to obtain mean shares will be via the contraction mapping above.

To sum up, Berry's (1994) two step GMM approach with a matrix of instruments $\mathbf{Z}$ is as follows:

Step 1 Compute $\widehat{\delta}_{jt}$.

- Without loss of generality, subsume p_{jt} within $\mathbf{x}_{jt}$ as just another column (a special attribute of product J), and rather than introduce new (unnecessary) notation, redefine $\mathbf{x}_{jt} = \begin{bmatrix}-p_{jt} & \mathbf{x}_{jt}\end{bmatrix}$. Similarly, redefine matrix $\mathbf{X}$ to be inclusive of the price vector so that $\mathbf{X} = \begin{bmatrix}\mathbf{p} & \mathbf{X}\end{bmatrix}$. Also, let $\mathbf{s}_t$ be the vector of observed shares and $\boldsymbol{\theta}_1 = \begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}'$.
- Conveniently, $\widehat{\delta}_{jt} = \ln(s_{jt}) - \ln(s_{0t})$ (in the case of simple logit) and $\widehat{\boldsymbol{\delta}} = \ln(\mathbf{s}) - \ln(\mathbf{s}_0)$
- Then $\xi_{jt}(\boldsymbol{\theta}_1) = \widehat{\delta}_{jt}(\mathbf{s}_t) - \mathbf{x}_{jt}\boldsymbol{\theta}_1$ – and in matrix notation, $\boldsymbol{\xi}(\boldsymbol{\theta}_1) = \widehat{\boldsymbol{\delta}} - \mathbf{X}\boldsymbol{\theta}_1$.

Step 2 Define the moment conditions as $\mathrm{E}(\mathbf{Z}'\boldsymbol{\xi}(\boldsymbol{\theta}_1)) = \mathbf{0}$.

- Next, $\min_{\boldsymbol{\theta}_1} \boldsymbol{\xi}(\boldsymbol{\theta}_1)'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\boldsymbol{\xi}(\boldsymbol{\theta}_1)$ where $\boldsymbol{\Phi} = (\mathrm{E}[\mathbf{Z}'\boldsymbol{\xi}\boldsymbol{\xi}'\mathbf{Z}])^{-1}$.
- In the case of logit, we have an analytical solution – see equation (5.37) in the GMM section, and replace $\mathbf{y}$ in that equation with $\widehat{\boldsymbol{\delta}}$:
  $\widehat{\boldsymbol{\theta}}_1 = (\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\widehat{\boldsymbol{\delta}}$
- Since we don't know $\boldsymbol{\Phi}$, we start with $\boldsymbol{\Phi} = \mathbf{I}$ or $\boldsymbol{\Phi} = (\mathbf{Z}'\mathbf{Z})^{-1}$, get an initial estimate of $\boldsymbol{\theta}_1$, use this to get residuals, and then recompute $\boldsymbol{\Phi} = (\mathrm{E}[\mathbf{Z}'\boldsymbol{\xi}\boldsymbol{\xi}'\mathbf{Z}])^{-1}$ to get the new estimates of $\boldsymbol{\theta}_1$.

6. Random Coefficients Logit/Heterogenous Tastes

We return now to the general set of utility for consumer n for product j in market t as in (5.1) but without imposing the restriction that taste parameters $\{\alpha, \boldsymbol{\beta}\}$ – the marginal utilities of product characteristics – are the same for all consumers. Thus, let the utility be given by

$$\begin{aligned} &u_{njt} = \alpha_n(y_n - p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta}_n + \xi_{jt} + \epsilon_{njt}, \text{ where} \\ &n = 1, \ldots, N, \qquad j = 0 \ldots, J, \qquad t = 1 \ldots, T. \end{aligned} \tag{6.1}$$

Each consumer is assumed to have a fixed set of coefficients $\{\alpha_n, \boldsymbol{\beta_n}\}$ but the person specific coefficients are modeled as a function of underlying common parameters $\{\boldsymbol{\Pi}$ and $\boldsymbol{\Sigma}\}$ that are multiplied to the person specific characteristics $(\boldsymbol{d}_n, \boldsymbol{\nu}_n)$, each of which are random draws from an underlying mean zero population with distribution functions $F_{\boldsymbol{d}}(\boldsymbol{d})$ and $F_{\boldsymbol{\nu}}(\boldsymbol{\nu})$. Thus,

$$\begin{aligned}\begin{bmatrix}\alpha_n\\ \boldsymbol{\beta}_n\end{bmatrix} &= \underbrace{\begin{bmatrix}\alpha\\ \boldsymbol{\beta}\end{bmatrix}}_{\boldsymbol{\theta}_1} + \underbrace{\boldsymbol{\Pi d}_n + \boldsymbol{\Sigma\nu}_n}_{\boldsymbol{\theta}_2=\{\boldsymbol{\Pi},\boldsymbol{\Sigma}\}} \\ &= \begin{bmatrix}\alpha\\ \boldsymbol{\beta}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\Pi}_\alpha\\ \boldsymbol{\Pi}_{\boldsymbol{\beta}}\end{bmatrix}\boldsymbol{d}_n + \begin{bmatrix}\boldsymbol{\Sigma}_\alpha\\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\nu}_{n\alpha} & \boldsymbol{\nu}_{\boldsymbol{n\beta}}\end{bmatrix}\end{aligned} \tag{6.2}$$

and where

$$\boldsymbol{d}_n \sim F_{\boldsymbol{d}}(\boldsymbol{d}) \qquad \boldsymbol{\nu}_n \sim F_{\boldsymbol{\nu}}(\boldsymbol{\nu}). \tag{6.3}$$

The person specific coefficients are equal to the <u>mean</u> value of the parameters $\boldsymbol{\theta}_1 = \begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}'$, plus deviation from the mean due to a second set of parameters $\boldsymbol{\theta}_2 = \{\boldsymbol{\Pi}, \boldsymbol{\Sigma}\}$ and given by $\boldsymbol{\Pi d}_n + \boldsymbol{\Sigma \nu}_n$.

To understand the dimensionality of these parameters, assume for concreteness that there are three observed product characteristics (so $k - 1 = 3$) and five observed person specific characteristics so that $\begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}'$ is a 4×1 vector (the additional dimension is for price) and $\boldsymbol{d}_n$ is a 5×1 vector. Additionally, $\boldsymbol{\nu}_n$ is also a 4×1 vector – these are the person specific random error terms that provide part of the deviation from the mean values of $\begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}'$. Then $\boldsymbol{\Pi}$ is 4×5 matrix (20 parameters) and $\boldsymbol{\Sigma}$ is a 4×4 matrix (16 parameters) and so the total number of parameters affecting the utility function are $4 + 20 + 16 = 40$. Let π_{ab} and σ_{ef} be the terms of $\boldsymbol{\Pi}$ and $\boldsymbol{\Sigma}$ respectively and let $(\boldsymbol{d}_n = (d_{1n}, \ldots, d_{5n})')$ be the five demographics of the n^{th} person recorded as deviation from the population mean values. Then

$$\begin{aligned}\alpha_n = \alpha \quad & +\pi_{11}d_{1n} + \pi_{12}d_{2n} + \ldots + \pi_{15}d_{5n} \\ & +\sigma_{11}v_{1n} + \sigma_{12}v_{2n} + \ldots + \sigma_{14}v_{4n} \\ \beta_{kn} = \beta_k \quad & +\pi_{k1}d_{1n} + \pi_{k2}d_{2n} + \ldots + \pi_{k5}d_{5n} \\ & +\sigma_{k1}v_{1n} + \sigma_{k2}v_{2n} + \ldots + \sigma_{k4}v_{4n},\end{aligned} \tag{6.4}$$

and if $\boldsymbol{d}_n$ are deviations from the mean and $\boldsymbol{\nu}_n$ are mean zero error terms, then $\mathrm{E}(\alpha_n) = \alpha$ and $\mathrm{E}(\boldsymbol{\beta}_n) = \boldsymbol{\beta}$. In general, if there are D person specific observed characteristics $(\boldsymbol{d}_n = (d_{1n}, \ldots, d_{Dn})')$ and $k-1$ product characteristics, then $\boldsymbol{\Pi}$ is a $k \times D$ and Σ is a $k \times k$ matrix of parameters, i.e.,

$$\underbrace{\begin{bmatrix}\alpha_n\\ \boldsymbol{\beta}_n\end{bmatrix}}_{k\times 1} = \underbrace{\begin{bmatrix}\alpha\\ \boldsymbol{\beta}\end{bmatrix}}_{k\times 1} + \underbrace{\boldsymbol{\Pi d}_n}_{k\times D \text{ by } D\times 1} + \underbrace{\boldsymbol{\Sigma\nu}_n}_{k\times k \text{ by } k\times 1}. \tag{6.5}$$

If we insert (6.2) back into (6.1) and simplify, then the utility function can be decomposed into three parts (or four, if we count $\alpha_n y_n$ term, but it drops out later on) and can be written as

$$\begin{aligned} &u_{njt} = \alpha_n y_n + \delta_{jt} + \mu_{njt} + \epsilon_{njt} \\ &\text{where,} \\ &\delta_{jt} = \delta(\mathbf{x}_{jt}, p_{jt}, \xi_{jt}; \boldsymbol{\theta}_1) = \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt} \\ &\mu_{njt} = \mu(\mathbf{x}_{jt}, p_{jt}, \boldsymbol{d}_n, \boldsymbol{\nu}_n; \boldsymbol{\theta}_2) = (-p_{jt}, \mathbf{x}_{jt})(\boldsymbol{\Pi}\boldsymbol{d}_n + \boldsymbol{\Sigma}\boldsymbol{\nu}_n). \end{aligned} \tag{6.6}$$

Note the similarity to the simple logit case: except for the μ_{njt} term, which arises due to multiplication of $(\boldsymbol{\Pi}\boldsymbol{d}_n + \boldsymbol{\Sigma}\boldsymbol{\nu}_n)$ with the observed product characteristics, the rest of the form is the same as in the logit case. As before, $\alpha_n y_n$ will drop out of the model, δ_{jt} is the mean utility of product j and is common to all consumers and $\mu_{njt} + \epsilon_{njt}$ is the mean-zero heteroscedastic error term that captures the deviation from the mean utility.

Recall that in the logit model the IIA property was arising due to the independence of the error terms ϵ_{njt}. The way around this problem is to allow these error terms to be correlated across different brands – and in principle one can allow a completely unrestricted variance-covariance matrix for the shocks ϵ_{njt}. This however leads back to the dimensionality problem as one has to estimate a large a large number of parameters (all pair-wise covariances between products and variances of each of the J products). The nested logit took a restricted version of this by imposing some structure on the error terms so that all products within a group have a correlation between them but not with those in other groups. In the current context, we retain the iid extreme value distribution assumption on ϵ_{njt}, but the correlation among the choices is generated via the μ_{njt} component of the composite error term $\mu_{njt} + \epsilon_{njt}$. As can be seen above from (6.6), correlation between utility of different products is a function of both product and consumer attributes so that products with similar characteristics will have similar rankings and consumers with similar demographics will have also have similar rankings of products ($\mu_{njt} = (-p_{jt}, \mathbf{x}_{jt})(\boldsymbol{\Pi}\boldsymbol{d}_n + \boldsymbol{\Sigma}\boldsymbol{\nu}_n)$). The advantage here is that rather than estimate a large number of parameters of a completely unrestricted variance-covariance matrix for ϵ_{njt}, we need to estimate relatively fewer parameters $\boldsymbol{\theta}_1 = (\alpha,\ \boldsymbol{\beta})', \boldsymbol{\theta}_2 = \{\boldsymbol{\Pi}, \boldsymbol{\Sigma}\}$.

In equation (6.6) the utility for product j for two different consumers differs only by $\mu_{njt} + \epsilon_{njt}$ (the δ_j term is the same for all consumers and $\alpha_n y_n$ is the same for all choices) and hence the fact that one consumer choose product j while another chooses product i must only be because the two consumers differ in their product specific idiosyncratic error terms $\mu_{njt} + \epsilon_{njt}$. Thus, as before, we can describe each consumer as a tuple of demographic and product specific shocks $(\boldsymbol{d}_n, \boldsymbol{\nu}_n, \epsilon_{n0t}, \epsilon_{n1t}, \ldots, \epsilon_{nJt})$, which implicity defines the set of individual attributes that choose product j given by

$$\mathbb{A}_{jt}(\mathbf{x}_t, \mathbf{p}_t, \boldsymbol{\delta}_t(\mathbf{x}_t, \mathbf{p}_t; \boldsymbol{\theta}_1); \boldsymbol{\theta}_2) = \{(\boldsymbol{d}_{nt}, \boldsymbol{\nu}_{nt}, \epsilon_{n0t}, \epsilon_{n1t}, \ldots, \epsilon_{nJt}) \mid u_{njt} > u_{nlt} \qquad \forall\, l = 0, 1, 2 \ldots J,\ l \neq j\}. \tag{6.7}$$

Note that this set is defined via the parameters $\boldsymbol{\theta}_2 = \{\boldsymbol{\Pi}, \boldsymbol{\Sigma}\}$ since they were part of the μ_{njt} term. As before, the market share of product j is the integral of the joint distribution of $(\boldsymbol{d}, \boldsymbol{\nu}, \boldsymbol{\epsilon})$ over the mass of individuals in the region A_{jt},

$$\begin{aligned} s_{jt} &= \int_{\mathbb{A}_{jt}} dF(\boldsymbol{d}, \boldsymbol{\nu}, \boldsymbol{\epsilon}) \\ &= \int_{\mathbb{A}_{jt}} dF_{\boldsymbol{d}}(\boldsymbol{d}) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}) dF_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \end{aligned} \tag{6.8}$$

where the second part follows only if we assume that the three random variables for a given consumer are independently distributed.

Unlike the logit case, the integral given above for the market share does not in general have a closed analytic form and hence has to evaluated using numerical methods. If however, we continue to assume that ϵ_{njt} has iid extreme value distribution, then the probability that a given individual $\tilde{n}$ – with endowed values of $\tilde{\boldsymbol{d}}_n$ and $\tilde{\boldsymbol{\nu}}_n$, or equivalently with a given value of $\tilde{\mu}_{njt}$ – chooses product j, continues to have a closed logit form like equation 5.6 and in this case is given by

$$s_{njt} = \frac{\exp(\delta_{jt} + \tilde{\mu}_{njt})}{\sum_{j=0}^{J} \exp(\delta_{jt} + \tilde{\mu}_{njt})}. \tag{6.9}$$

Since $\mu_{njt} = \mu(\mathbf{x}_{jt}, p_{jt}, \boldsymbol{d}_n, \boldsymbol{\nu}_n; \boldsymbol{\theta}_2)$ integrating this individual probability over the distribution of $\boldsymbol{d}_n$ and $\boldsymbol{\nu}_n$ will recover the market share of product j and hence equation (6.8) becomes

$$\begin{aligned} s_{jt} &= \int_{\mathbb{A}_{jt}} s_{njt} dF_{\boldsymbol{d}}(\boldsymbol{d}) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}) \\ &= \int_{\mathbb{A}_{jt}} \Big\{ \frac{\exp(\delta_{jt} + \mu_{njt})}{\sum_{j=0}^{J} \exp(\delta_{jt} + \mu_{njt})} \Big\} dF_{\boldsymbol{d}}(\boldsymbol{d}) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}). \end{aligned} \tag{6.10}$$

In this case, the price elasticities of market shares are given by

$$\eta_{jkt} = \frac{\partial s_{jt}}{\partial p_{kt}} \frac{p_{kt}}{s_{jt}} = \begin{cases} -\frac{p_{jt}}{s_{jt}} \int_{\mathbb{A}_{jt}} \alpha_n s_{njt}(1 - s_{njt}) dF_{\boldsymbol{d}}(\boldsymbol{d}) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}) & \text{if } j = k, \\ \frac{p_{kt}}{s_{jt}} \int_{\mathbb{A}_{jt}} \alpha_n s_{njt} s_{nkt} dF_{\boldsymbol{d}}(\boldsymbol{d}) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}) & \text{otherwise.} \end{cases} \tag{6.11}$$

The main advantages of this model are that estimation requires estimation of a handful of parameters (rather than square of the number of parameters), elasticities do not exhibit the problems noted earlier for the logit (own or cross-elasticities) and allows us to model consumer heterogeneity rather than rely on a representative consumer. Nonetheless, nothing comes for free as this model is clearly harder to estimate and requires some prior knowledge of the distributions of $\boldsymbol{d}_n$ and $\boldsymbol{\nu}_n$ and we need to evaluate the integral numerically.

Integration. To compute the integral in (6.10), we need to know the distribution functions $F_{\boldsymbol{d}}(\boldsymbol{d})$ and $F_{\boldsymbol{\nu}}(\boldsymbol{\nu})$. Unless these two distributions take some explicitly convenient form, so that the whole expression reduces to a closed form analytical formula, integration will need to be done numerically. Since $\boldsymbol{d}_n$ is the vector of demographics for consumer n (income, family size, age, gender, etc.), one way to proceed is to make use other of data sources, such as the census data, to construct a non-parametric distribution. We can then take random draws from this distribution to compute the integral above. In practice one can directly draw N number of consumers – where N is a reasonably large number – from each of the t markets and record their demographic information. Thus, let us assume that $\boldsymbol{d}_n$ is a 5×1 vector of demographics, and that we have obtained N_s random draws from each market and recorded the values of these demographics. Next, recall that if $\mathbf{x}_{jt}$ is a vector of three observed characteristics $(k-1=3)$ for product j, then for each person, $\boldsymbol{\nu}_n$ is a 4×1 (or more generally $k \times 1$) vector of random error terms that provide part of the deviation from the mean values of $\begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}'$. To this end, most researchers often specify $F_{\boldsymbol{\nu}}(\boldsymbol{\nu})$ as standard multivariate normal and take N draws per market to obtain $\boldsymbol{\nu}_n$. Let us again assume that with the help of a good random number generater, we have taken N_s such draws per market and have recorded a series of 4×1 vectors for each person.

The key to understanding numeric integration by simulation is as follows. Think of any arbitrary random variable x (nothing to do with characteristic vector $\mathbf{x}_{jt}$ above) with a probability distribution $f(x) = dF(x)/dx \rightarrow dF(x) = f(x)dx$. If we compute the integral $\int x \cdot f(x)dx$, this is just the expected value of x, i.e., $\mathrm{E}[x] = \int x \cdot dF(x)$, and the sample analog would be the weighted average of x given by $\bar{x} = \sum_n x_n Pr(x_n)$. If all values are equally possible, then it is just the simple average $\bar{x} = (1/N)\sum_n x_n$. The idea carries over to any function $g(x)$ defined over x such that $\mathrm{E}[g(x)] = \int g(x) \cdot dF(x)$, and the sample analog would be $\overline{g(x)} = \sum_n g(x_n)Pr(x_n)$. Thus, if we wanted to numerically evaluate the integral of $g(x)$ with a known distribution of x (i.e., evaluate $\int g(x) \cdot dF(x)$), all we need to do is take lots of draws of x from this known distribution, evaluate $g(x)$ at each of these points and then just take a simple average of all these values of $g(x)$. We will get a pretty good value of the integral by this method if we have taken enough *good* draws of the random variable x. Consider the case where x is distributed between 0 and 3 such that the probabilities of draws are $\Pr(0 \leq x < 1) = .45$, $\Pr(1 \leq x < 2) = .1$, and $\Pr(2 \leq x < 3) = .45$. If we drew 100 random numbers from this distribution, we would expect about 45 of them to be between 0 and 1, another 10 observations between 1 and 2, and 45 observations between 2 and 3. If that were the case, we could safely evaluate $g(x)$ at each of these 100 random draws and take their average to compute $\mathrm{E}[g(x)] = \int g(x) \cdot dF(x)$. If on the other hand we find that the drawing sequence (algorithm) is such that for the first 100 draws, we have 1/3 of observations from each of the three regions, then with just 100 draws, average values of $g(x)$ will obviously give a very poor (if not outright wrong) approximation to the integral in question. (There is a large literature on drawing from different types of random distributions, for a good review of basic techniques see chapter 9 in Train).

Thus given the values of the parameters $\boldsymbol{\theta}_2 = \{\boldsymbol{\Pi}, \boldsymbol{\Sigma}\}$, a value of mean utility δ_{jt} and N_s random values of $\boldsymbol{d}_n$ and $\boldsymbol{\nu}_n$, the predicted market share of good j can be computed using the smooth simulator as the average value of s_{njt} over the N_s observations,

$$\begin{aligned}\tilde{s}_{jt} &= \int_{A_{jt}} s_{njt} dF_{\boldsymbol{d}}(\boldsymbol{d}) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}) \\ &= \frac{1}{N_s}\sum_{n}^{N_s} s_{njt} = \frac{1}{N_s}\sum_{n}^{N_s}\left\{\frac{\exp(\delta_{jt}+\mu_{njt})}{\sum_{j=0}^{J}\exp(\delta_{jt}+\mu_{njt})}\right\} \\ &\text{where } \mu_{njt} = (-p_{jt}, \mathbf{x}_{jt})(\boldsymbol{\Pi}\boldsymbol{d}_n + \boldsymbol{\Sigma}\boldsymbol{\nu}_n).\end{aligned} \tag{6.12}$$

Distributions of $\boldsymbol{\nu}_n$ and Parameters $\boldsymbol{\theta}_2$. The number of parameters to estimate is still somewhat large – in the simple case of five demographics and 3+1 product characteristics, we saw that the total number of parameters is 40 – and data may not allow very precise measurement of all these parameters. In BLP, they do not use individual demographics to create variation in person specific coefficients. Equivalently, the $k \times d$ matrix $\boldsymbol{\Pi}$ consists of zeros and the variation in $\begin{bmatrix}\alpha_n & \boldsymbol{\beta}_n'\end{bmatrix}'$ is only due to $\boldsymbol{\Sigma}\boldsymbol{\nu}_n$. By contrast, Nevo sets only some of the terms of $\boldsymbol{\Pi}$ to zero and estimates the other coefficients. More often however, researchers set $\boldsymbol{\Sigma}$ as a diagonal matrix and estimate only the leading terms of this matrix. This is not as restrictive as it may appear at first pass.

To understand the logic of choosing parameters that are set to zero, and the implications, let us consider a very simple example (only to keep the algebra manageable) where there is only one observed characteristic of each product, plus price, so that $\begin{bmatrix}\alpha_n & \boldsymbol{\beta}_n'\end{bmatrix}'$ is just a 2×1 column vector instead of $k \times 1$ (just to be clear, in what follows in the next couple of paragraphs, think of $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}$ as just 1×1 scalars even though I will continue to write them in bold font for vectors). Further, to understand the implications/logic of setting the off-diagonals of $\boldsymbol{\Sigma}$ to zero, let's further simplify and suppose that all the elements of $\boldsymbol{\Pi}$ are zero (again, only to simplify the algebra as the main idea carries through with or without $\boldsymbol{\Pi}$ in the utility function).

Thus sans the $\boldsymbol{\Pi}\boldsymbol{d}_n$ term, let

$$\begin{bmatrix}\alpha_n \\ \boldsymbol{\beta}_n\end{bmatrix} = \begin{bmatrix}\alpha \\ \boldsymbol{\beta}\end{bmatrix} + \boldsymbol{\Sigma}\boldsymbol{\nu}_n = \begin{bmatrix}\alpha \\ \boldsymbol{\beta}\end{bmatrix} + \begin{bmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22}\end{bmatrix}\begin{bmatrix}\nu_{1n} \\ \nu_{2n}\end{bmatrix}, \tag{6.13}$$

and compute the expected value and variance of α_n and $\boldsymbol{\beta}_n$. Since $\boldsymbol{\nu}_n$ is a mean zero error term, then

$$\begin{aligned}
&\alpha_n = \alpha + \sigma_{11}\nu_{1n} + \sigma_{12}\nu_{2n} \\
&\boldsymbol{\beta}_n = \boldsymbol{\beta} + \sigma_{21}\nu_{1n} + \sigma_{22}\nu_{2n} \\
&\mathrm{E}[\alpha_n] = \alpha \\
&\mathrm{E}[\boldsymbol{\beta}_n] = \boldsymbol{\beta} \\
&\mathrm{Var}[\alpha_n] = \sigma_{11}^2\mathrm{Var}[\nu_{1n}] + 2\sigma_{11}\sigma_{12}\mathrm{Cov}[\nu_{1n}, \nu_{2n}] + \sigma_{12}^2\mathrm{Var}[\nu_{2n}] \\
&\mathrm{Var}[\boldsymbol{\beta_n}] = \sigma_{21}^2\mathrm{Var}[\nu_{1n}] + 2\sigma_{21}\sigma_{22}\mathrm{Cov}[\nu_{1n}, \nu_{2n}] + \sigma_{22}^2\mathrm{Var}[\nu_{2n}].
\end{aligned} \tag{6.14}$$

The implications of setting the off-diagonal terms in $\boldsymbol{\Sigma}$ to zero are now obvious: if $\sigma_{12} = \sigma_{21} = 0$, then α_n is a deviation from the mean value of α and the deviation is determined only by a random shock ν_{1n} multiplied by a coefficient σ_{11}, but the shock to the marginal utility of the second characteristic ν_{2n}, does not affect the deviation from mean for the first characteristics, i.e., the marginal (dis)utility of price. Put another way, the unobserved heterogeneity has been modeled such that if price and speed of a computer are the only two characteristics in consideration, and a given person gets a positive shock to the marginal utility of speed (i.e., they get more utility from the speed of computer relative to another person), then it does not imply that they also get a higher (dis)utility from the price of the computer due to the higher utility from speed. The (dis)utility from price is equal to α plus a person specific deviation only for price $\sigma_{11}\nu_{1n}$. Similarly, marginal utility from speed $\boldsymbol{\beta}_n$ is a deviation from $\boldsymbol{\beta}$ and only depends on $\sigma_{22}\nu_{2n}$ but not on ν_{1n}.

Similarly, variances of α_n and $\boldsymbol{\beta}_n$ depend on the variances of the shocks for these characteristics (e.g. $\mathrm{Var}[\alpha_n] = \sigma_{11}^2\mathrm{Var}[\nu_{1n}]$) but not on the *covariance* of the shocks, even if $\mathrm{Cov}[\nu_{1n}, \nu_{2n}] \neq 0$, since $\sigma_{12} = \sigma_{21} = 0$. Carrying on with this simple case where $\boldsymbol{\beta}_n$ is a scaler, let's also compute the covariance between α_n and $\boldsymbol{\beta}_n$. Covariance between the two random variables is defined as $\mathrm{Cov}(\alpha_n, \boldsymbol{\beta}_n) = \mathrm{E}\left[\{\alpha_n - \mathrm{E}(\alpha_n)\}\{\boldsymbol{\beta}_n - \mathrm{E}(\boldsymbol{\beta}_n)\}\right]$, and hence,

$$\begin{aligned}
\mathrm{Cov}(\alpha_n, \boldsymbol{\beta}_n) &= \mathrm{E}(\alpha_n\boldsymbol{\beta}_n) - \alpha\boldsymbol{\beta} \\
&= \sigma_{11}\sigma_{21}\mathrm{Var}(\nu_{1n}) + \sigma_{12}\sigma_{22}\mathrm{Var}(\nu_{2n}) \\
&+ \sigma_{11}\sigma_{22}\mathrm{Cov}(\nu_{1n}, \nu_{2n}) + \sigma_{12}\sigma_{21}\mathrm{Cov}(\nu_{1n}, \nu_{2n}) \\
&= \sigma_{11}\sigma_{22}\mathrm{Cov}(\nu_{1n}, \nu_{2n}).
\end{aligned} \tag{6.15}$$

In the equation above, the first line is due to the definition of a covariance and the observation that $\mathrm{E}[\alpha_n] = \alpha$ and $\mathrm{E}[\boldsymbol{\beta}_n] = \boldsymbol{\beta}$, and the second line follows from substituting values of α_n and $\boldsymbol{\beta}_n$ from equation (6.14), taking the expectations, setting $\mathrm{E}[\boldsymbol{\nu}_n] = \mathbf{0}$ and simplifying. The last line is if we set $\sigma_{12} = \sigma_{21} = 0$ and shows that even after setting the off-diagonals in $\boldsymbol{\Sigma}$ equal to zero, the covariance between the marginal utilities is not necessarily zero – *unless we now further assume that the mean zero error terms* $\boldsymbol{\nu}_n$ *are not correlated across the characteristics.*

As it turns out, it is also common to assume that $\boldsymbol{\nu}_n$ are drawn from multivariate standard normal or log normal, i.e., covariances between the error terms are zero as well. In the special case where the terms of $\mathbf{\Pi}$ are also zero – as in the foregoing discussion – this implies that covariances between marginal utilities will also be zero. However, if terms of $\mathbf{\Pi}$ are not all zero, they will still invoke correlations between the marginal utilities of different characteristics as equation (6.4), reproduced below for this special case of two characteristics and five demographics, shows

$$\begin{aligned} \alpha_n = \alpha \quad & +\pi_{11} d_{1n} + \pi_{12} d_{2n} + \ldots + \pi_{15} d_{5n} \\ & \qquad\qquad\qquad\qquad +\sigma_{11}\nu_{1n} + \sigma_{12}\nu_{2n} \\ \boldsymbol{\beta}_n = \boldsymbol{\beta} \quad & +\pi_{21} d_{1n} + \pi_{22} d_{2n} + \ldots + \pi_{25} d_{5n} \\ & \qquad\qquad\qquad\qquad +\sigma_{21}\nu_{1n} + \sigma_{22}\nu_{2n}. \end{aligned} \tag{6.4}$$

In this case, the covariance between α_n and $\boldsymbol{\beta_n}$ will be invoked via the π terms and the covariances between the demographic variables, even if we set $\sigma_{12} = \sigma_{21} = 0$ and choose the distribution of $\boldsymbol{\nu}_n$ to be multivariate standard normal. Thus as mentioned earlier, if we use demographic data and don't set the $\mathbf{\Pi}$ to zero (at least not all terms) then setting the off diagonals of $\mathbf{\Sigma}$ to zero and drawing $\boldsymbol{\nu}_n$ from multivariate standard normal is not so restrictive.

6.1. Estimation Details

We finally turn to the estimation of the random coefficients model. The essential idea of estimation remains the same as that of two-step estimation outlined in section 5.5. I recommend re-reading the summary of the Berry's two step method outlined for the case of logits and the points of departure for more general cases before starting this section. Briefly, estimate mean utility δ_{jt} and then use it in the second step to estimate the moment functions and find parameters that minimize the value. This of course requires first estimating model predicted market shares via (6.10), equating them to observed market shares, and then inverting the relation and using a contraction mapping to compute δ_{jt}. We consider each of these along the way and following Nevo (2001), combine everything in a 5-step algorithm.

(-1) For each market t, draw N_s random values for $(\boldsymbol{\nu}_n, \boldsymbol{d}_n)$ from the distributions $F_{\boldsymbol{\nu}}(\boldsymbol{\nu})$ and $F_{\boldsymbol{d}}(\boldsymbol{d})$. The distribution $F_{\boldsymbol{d}}(\boldsymbol{d})$ can be estimated using census data. For $F_{\boldsymbol{\nu}}(\boldsymbol{\nu})$ we can use zero mean multivariate normal with a pre-specified covariance matrix.

(0) Select arbitrary initial values of δ_{jt} and $\boldsymbol{\theta}_2 = \{\mathbf{\Pi}, \mathbf{\Sigma}\}$. For $\boldsymbol{\theta}_1 = \begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}'$ use values from simple logit estimation.

(1) Use random draws and the initial parameter values to estimate model predicted market shares $\tilde{s}_{jt}$ of each product in each market (use (6.12)).

(2) Obtain $\widehat{\delta}_{jt}$.

(a) Keep $\boldsymbol{\theta}_2 = \{\boldsymbol{\Pi}, \boldsymbol{\Sigma}\}$ fixed and change values of δ_{jt} until predicted shares $\tilde{s}_{jt}$ in step above, equal the observed shares. This is the inversion step where we want to find $\boldsymbol{\delta}_t$ such that $s_{jt} = \tilde{s}_{jt}(\delta_{1t}, \ldots, \delta_{Jt}, \boldsymbol{\theta}_2)$ in each market.

(b) This can be done using the contraction mapping $\boldsymbol{\delta}_t^{h+1} = \boldsymbol{\delta}_t^h + [ln(\mathbf{s}_t) - ln(\tilde{\mathbf{s}}_t)]$.

(c) Note carefully that mean utility is a function of observed market shares and parameters $\boldsymbol{\theta}_2$. Thus, $\delta_{jt} = \delta_{jt}(\mathbf{s}_t, \boldsymbol{\theta}_2)$.

(3) Define error term as $\xi_{jt} = \widehat{\delta}_{jt}(\mathbf{s}_t, \boldsymbol{\theta}_2) + \alpha p_{jt} - \mathbf{x}_{jt}\boldsymbol{\beta}$ and calculate the value of the moment condition, i.e., the GMM objective function.

(a) As before, subsume p_{jt} within $\mathbf{x}_{jt}$ as just another column of $\mathbf{x}_{jt}$ and redefine $\mathbf{x}_{jt} = \begin{bmatrix} -p_{jt} & \mathbf{x}_{jt} \end{bmatrix}$. Similarly, redefine matrix $\mathbf{X}$ to be inclusive of the price vector so that $\mathbf{X} = \begin{bmatrix} -\mathbf{p} & \mathbf{X} \end{bmatrix}$.

(b) Thus $\xi_{jt}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \widehat{\delta}_{jt}(\mathbf{s}_t, \boldsymbol{\theta}_2) - \mathbf{x}_{jt}\boldsymbol{\theta}_1$.
In matrix notation $\boldsymbol{\xi} = \widehat{\boldsymbol{\delta}}(\mathbf{s}, \boldsymbol{\theta}_2) - \mathbf{X}\boldsymbol{\theta}_1$.

(c) Then the objective function to be minimized is
$\Big(\boldsymbol{\xi}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'\mathbf{Z}\Big)\boldsymbol{\Phi}\Big(\mathbf{Z}'\boldsymbol{\xi}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\Big)$,
where $\boldsymbol{\Phi}$ is the GMM weighting matrix.

(d) Initially set the weighting matrix as $\boldsymbol{\Phi} = (\mathbf{Z}'\mathbf{Z})^{-1}$.

(4) Search for better values of $\boldsymbol{\theta}_1 = \begin{bmatrix} \alpha & \boldsymbol{\beta}' \end{bmatrix}'$ and $\boldsymbol{\theta}_2 = \{\boldsymbol{\Pi}, \boldsymbol{\Sigma}\}$ and the GMM weighting matrix Φ as follows:

(a) Before you start searching for the parameter values that minimize the objective function, note that while $\boldsymbol{\xi}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is a function of both sets of parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, it actually partitions into two components: $\xi_{jt}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \widehat{\delta}_{jt}(\mathbf{s}_t, \boldsymbol{\theta}_2) - \mathbf{x}_{jt}\boldsymbol{\theta}_1$. This is important because we can help the search algorithm by solving for $\boldsymbol{\theta}_1$, conditional on $\boldsymbol{\theta}_2$ analytically. How? In the GMM objective function given above $[(\boldsymbol{\xi}'\mathbf{Z})\boldsymbol{\Phi}(\mathbf{Z}'\boldsymbol{\xi})]$, set $\boldsymbol{\xi} = \widehat{\boldsymbol{\delta}}(\boldsymbol{\theta}_2) - \mathbf{X}\boldsymbol{\theta}_1$ (I have suppressed the dependence on observed shares to keep it simple). Now consider the first order condition with respect to $\boldsymbol{\theta}_1$ and solve for $\boldsymbol{\theta}_1$. See equations 5.31 and 5.32 for FOC and its solution for the GMM estimator. This implies that if we have some fixed values of $\boldsymbol{\theta}_2$, then $\boldsymbol{\theta}_1$ can be solved for analytically as $\boldsymbol{\theta}_1 = (\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\widehat{\boldsymbol{\delta}}(\boldsymbol{\theta}_2)$.

(b) Thus, first solve (search) for $\boldsymbol{\theta}_1$ as $\widehat{\boldsymbol{\theta}}_1 = (\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}'\widehat{\boldsymbol{\delta}}(\boldsymbol{\theta}_2)$.

(c) Use new $\boldsymbol{\theta}_1 = \begin{bmatrix} \alpha & \boldsymbol{\beta}' \end{bmatrix}'$ to re-compute error term $\boldsymbol{\xi}$ (see 3b above).

(d) Next, update the weighting matrix $\boldsymbol{\Phi}$ as $\boldsymbol{\Phi} = (\mathbf{Z}'\boldsymbol{\xi}\boldsymbol{\xi}'\mathbf{Z})^{-1}$.

(e) Take the new value of $\boldsymbol{\Phi}$ and update the GMM objective function, $(\boldsymbol{\xi}'\mathbf{Z})\boldsymbol{\Phi}(\mathbf{Z}'\boldsymbol{\xi})$.

(f) Finally, update $\boldsymbol{\theta}_2 = \{\boldsymbol{\Pi}, \boldsymbol{\Sigma}\}$ – do a non-linear search over $\{\boldsymbol{\Pi}, \boldsymbol{\Sigma}\}$ to minimize the objective function.

(5) Return to step (1) above with all new shiny parameter values (keep the original draws) and iterate. Note that you can skip the updating of the weighting matrix $\mathbf{\Phi}$ in step 4e from now on.

Some Further Details.

- **Brand Dummies**. In the section on logits, we discussed adding in the brand dummies to the vector $\mathbf{x}_{jt}$ and recovering the $\boldsymbol{\beta}$ coefficients for the brand characteristics. Same can be done here as well, but will need to have two separate versions of data matrix $\mathbf{X}$ (call them $\mathbf{X}_1$ and $\mathbf{X}_2$). Observe that $\mathbf{X}$ (defined to be inclusive of the price vector) enters the utility function twice: in the linear part of the estimation as mean utility $\boldsymbol{\delta}(\mathbf{X};\boldsymbol{\theta}_1) = \mathbf{X}\boldsymbol{\theta}_1 + \boldsymbol{\xi}$ – this is from $\delta_{jt} = \delta(\mathbf{x}_{jt}, p_{jt}, \xi_{jt}; \boldsymbol{\theta}_1) = \alpha(-p_{jt}) + \mathbf{x}_{jt}\boldsymbol{\beta} + \xi_{jt}$ – and in the non-linear part of the estimation as individual deviation from the mean utility $\boldsymbol{\mu}_n(\mathbf{X};\boldsymbol{\theta}_2, \boldsymbol{d}_n, \boldsymbol{\nu}_n) = \mathbf{X}(\mathbf{\Pi}\boldsymbol{d}_n + \mathbf{\Sigma}\boldsymbol{\nu}_n)$ – this follows from $\mu_{njt} = (-p_{jt}, \mathbf{x}_{jt})(\mathbf{\Pi}\boldsymbol{d}_n + \mathbf{\Sigma}\boldsymbol{\nu}_n)$ – and allows for random coefficients on product characteristics. In practice we may not want to allow random coefficients on all characteristics, in which case the data matrix $\mathbf{X}$ appearing in $\boldsymbol{\mu}_n$ can be a subset of the one appearing the linear part $\boldsymbol{\delta}$. Thus, we can write the two components as $\boldsymbol{\delta}(\mathbf{X}_1;\boldsymbol{\theta}_1) = \mathbf{X}_1\boldsymbol{\theta}_1 + \boldsymbol{\xi}$ and $\boldsymbol{\mu}_n(\mathbf{X}_2;\boldsymbol{\theta}_2, \boldsymbol{d}_n, \boldsymbol{\nu}_n) = \mathbf{X}_2(\mathbf{\Pi}\boldsymbol{d}_n + \mathbf{\Sigma}\boldsymbol{\nu}_n)$.

  In general then, $\mathbf{X}_1$ includes all variables that are common to all individuals (price, promotional activities, and brand characteristics <u>or brand dummies</u> instead of brand characteristics), while $\mathbf{X}_2$ contains variables that can have random coefficients (price and product characteristics but not brand dummies). Finally, note that if we use $\mathbf{X}_1$ and $\mathbf{X}_2$, then the estimator $\widehat{\boldsymbol{\theta}}_1$ in step 4a/4b above will be $\widehat{\boldsymbol{\theta}}_1 = (\mathbf{X}_1'\mathbf{Z}\mathbf{\Phi}\mathbf{Z}'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Z}\mathbf{\Phi}\mathbf{Z}'\widehat{\boldsymbol{\delta}}(\boldsymbol{\theta}_2)$.
- **Additional Instruments.** The instruments matrix $\mathbf{Z}$ consists of all exogenous variables. If the brand characteristics (excluding price) are exogenous, then the brand characteristics plus the instrument(s) for the price variable consist of the matrix $\mathbf{Z}$, or alternatively, if we use brand dummies, then the brand dummies and the price instrument(s) form the matrix $\mathbf{Z}$. However, note that if we have only one additional instrument for price, it will not be enough for identification of the model parameters. The brand characteristics (or brand dummies) plus the one additional instrument for price will give *exactly* as many moment conditions as the number of components of the parameter vector $\boldsymbol{\theta}_1$. These would be enough in the linear logit case. However, in the random coefficients case, we have to estimate additional $k \times D + k \times k$ parameters of $\boldsymbol{\theta}_2 = \{\mathbf{\Pi}, \mathbf{\Sigma}\}$. This is not possible unless we have additional $k \times D + k \times k$ moment conditions. In practice, researchers often set some of the terms of the $\mathbf{\Pi}$ matrix to zero (based on prior beliefs about the random coefficients for some of the product characteristics being due to differences in individual demographics) and also set the parameter matrix $\mathbf{\Sigma}$ to be diagonal (see earlier discussions). This reduces the need for

additional moment conditions from $kD + k^2$ to $g + k$ where g is the number of non-zero terms in $\mathbf{\Pi}$. These may be relatively easier to overcome (these instruments should also not be nearly collinear else will give rise to redundant moment conditions). For instance, if one is using BLP style instruments for price (and product characteristics are exogenous) then recall that, in general, one gets more than one instrument for price by using sums of the values of characteristics of other products offered by a firm, and the sums of the values of the same characteristics of products offered by other firms. Alternatively, if using Hausman style instruments, the price of the product from more than one market needs to be used (for instance, Nevo (2001) uses data from 20 quarters and multiple cities and constructs 20 additional instruments from other cities matching one from each quarter). An additional set of instruments could be the average value (average over n individuals) of the product characteristics interacted with the person specific demographics to account for the parameters in the $\mathbf{\Pi}$ matrix and similarly the average value of the person specific shocks $\boldsymbol{\nu}$ interacted with product characteristics.

7. Summary

These lecture notes are meant to be an aid in understanding basic estimation issues. They are by no means complete in the sense of covering all the important variants of the models discussed above. For instance, some useful and important variations to the random coefficients model discussed above include using individual level data (in addition to the aggregate data), adding in the cost side moment restrictions to the model (e.g. equation (3.15) $\mathbf{p} = \mathbf{c} + \mathbf{\Omega^{-1}q}(\mathbf{p}, \mathbf{z}; \boldsymbol{\xi})$) and modeling dynamic demand. Nonetheless, these lecture notes should serve as a useful starting point in understanding these variants. Finally, note that while canned routines in almost any software (SAS, STATA, etc.) can be used for the linear models (multi-budgeting with AIDs, logit, nested logit etc.) no canned routine (yet) exists for random coefficients models with aggregate data. Thus, while the step-by-step algorithm outlined in the previous section should help in coding for your own research project, some nearly canned routines in MATLAB, R, GAUSS etc. have been helpfully provided by several researchers (Aviv Nevo, K. Sudhir, and Matthijs Wildenbeest to name a few) and can serve as a good starting point for coding your own work.

Farasat A.S. Bokhari
School of Economics & Centre for Competition Policy
University of East Anglia

References

Ackerberg, D., Benkard, C. L., Berry, S., and Pakes, A. (2007). Econometric tools for analyzing market outcomes. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6A, chapter 63, page 41714276. Elsevier.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics*, 25(2):242–262.

Bokhari, F. A. S. and Fournier, G. M. (2013). Entry in the ADHD drugs market: Welfare impact of generics and me-toos. *Journal of Industrial Economics*, 61(2):340–393.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge.

Deaton, A. and Muellbauer, J. (1980). *Economics and consumer behavior*. Cambridge University Press, Cambridge, UK.

Hausman, J. A., Leonard, G., and Zona, J. (1994). Competitive analysis with differentiated products. *Annales d'Economie et de Statistique*, 34:159–180.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics and Management Strategy*, 9(4):513–548.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.