Notes on "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand" (Nevo, 2000)

- Paper: Nevo, Aviv, "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand," *Journal of Economics & Management Strategy*, 9 (2000), 513–548. https://doi.org/10.1111/j.1430-9134.2000.00513.x
- The Appendix of this paper is very useful as it provides details about the estimation steps. A copy can be found at https://www.rasmusen.org/zg601/readings/Nevo.Ras_guide_appendix.pdf

1. Introduction

- This paper describes an important development in methods for estimating random-coefficients discrete-choice models of demand (Berry et al., 1995).
- The method proposed by BLP maintains the advantage of the logit model in handling a large number of products. It is superior to prior methods because (1) the model can be estimated using only market-level price and quantity data, (2) it deals with the endogeneity of prices, and (3) it produces demand elasticities that are more realistic—for example, cross-price elasticities are larger for products that are closer together in terms of their characteristics.

2. The Model

2.1. Setup

- T markets, indexed by $t = 1, \cdots, T$.
- I_t consumers in market t, indexed by $i = 1, \cdots, I_t$.
- Aggregate quantity, average prices and product characteristics of J goods are observed, indexed by $j = 1, \cdots, J$.
- Indirect utility:

  $$U\left(x_{jt}, \xi_{jt}, p_{jt}, \tau_i; \theta\right),$$

  where x_{ji} denotes product-level observable attributes, ξ_{jt} denotes product-level unobservable attributes, p_{jt} denotes prices, τ_i denotes individual-level characteristics, and θ is parameters that need to be estimated with data.
  - Setting a linear specification for product-level attributes:

    $$u_{ijt} = \alpha_i \left(y_i - p_{jt}\right) + x_{jt}'\beta_i + \xi_{jt} + \varepsilon_{ijt}, \tag{1}$$

    where ε_{ijt} is a zero-mean stochastic term.
  - Individual-specific taste coefficients are modeled as a linear function of individual-level characteristics:

    $$\begin{pmatrix}\alpha_i \\ \beta_i\end{pmatrix} = \begin{pmatrix}\alpha \\ \beta\end{pmatrix} + \Pi D_i + \Sigma \nu_i, \tag{2}$$

    where D_i denotes observable (distribution of) characteristics of individuals and ν_i denotes unobservable characteristics of individuals. $\nu_i \sim P_\nu^*(\nu)$ is an assumed parametric distribution and $D_i \sim \hat{P}_D^*(D)$ is a known (non-)parametric distribution from data or having been estimated

elsewhere. For simplicity, assume that ν_i and D_i are independent. Since some components of these coefficients are unknown and are treated as random, it forms a random-coefficients model.

- Introducing an *outside good*, indexed by $j = 0$, whose utility is

$$u_{i0t} = \alpha_i y_i + \xi_{0t} + \pi_0 D_i + \sigma_0 \nu_{i0} + \varepsilon_{i0t}.$$

  Firstly, note that $\alpha_i y_i$ is constant given the decision maker, thus it is not identified and we can set $\alpha_i y_i = 0$. Moreover, ξ_{jt} is alternative-specific constant, therefore, among the $J + 1$ goods (including the outside good), only J constants can be identified and thereby we can set $\xi_{0t} = 0$. Furthermore, $\pi_j D_i$ reflects heterogeneous effects of the same observed individual characteristics vector D_i on the utility of different goods, thus we can identify only J effects of them. We can set $\pi_0 = 0$. Finally, $\sigma_0 \nu_{i0} + \varepsilon_{i0t}$ is the unobserved term. As only the differences between error terms matter, it is also safe to set $\sigma_0 \nu_{i0} + \varepsilon_{i0t} = 0$. Taken together, we set the utility of the outside good, u_{i0t}, as zero. In other words, we normalize the utility of each good based on the outside good. We should keep this in mind because the normalization influences the interpretation of the parameters to be estimated.
- Define $\theta \equiv (\theta_1, \theta_2)$ that contains all the parameters of the model, wherein $\theta_1 = (\alpha, \beta)$ contains K linear parameters and $\theta_2 = (\Pi, \Sigma)$ contains non-linear parameters. Therefore, the utility can be decomposed as

$$u_{ijt} = \alpha_i y_i + \delta_{jt}\left(x_{jt}, p_{jt}, \xi_{jt}; \theta_1\right) + \mu_{ijt}\left(x_{jt}, p_{jt}, \nu_i, D_i; \theta_2\right) + \varepsilon_{ijt}, \tag{3}$$

  where $\delta_{jt} \equiv x_{jt}'\beta - \alpha p_{jt} + \xi_{jt}$ is called "mean utility" and $\mu_{ijt} \equiv [-p_{jt}, x_{jt}'](\Pi D_i + \Sigma v_i)$ refers to the zero-mean heteroskedasticity deviation from the mean utility due to random coefficients.
- Define set A_{jt} as the individuals who choose good j in market t:

$$A_{jt}\left(x_{\cdot t}, p_{\cdot t}, \delta_{\cdot t}; \theta_2\right) = \{(D_i, \nu_i, \varepsilon_{i0t}, \cdots, \varepsilon_{iJt}) | u_{ijt} \geq u_{ilt},\ \forall l = 0, 1, \cdots, J\},$$

  where $x_{\cdot t}$ contains all attribute vector of goods in market t, and $p_{\cdot t}$ and $\delta_{\cdot t}$ take similar meanings. Accordingly, market share of the j-th good is:

$$\begin{aligned} s_{jt}\left(x_{\cdot t}, p_{\cdot t}, \delta_{\cdot t}; \theta_2\right) &= \int_{A_{jt}} \mathrm{d}P^*\left(D, \nu, \varepsilon\right) \\ &= \int_{A_{jt}} \mathrm{d}P^*\left(\varepsilon | D, \nu\right) \mathrm{d}P^*\left(\nu | D\right) \mathrm{d}P_D^*\left(D\right) \\ &= \int_{A_{jt}} \mathrm{d}P^*\left(\varepsilon\right) \mathrm{d}P^*\left(\nu\right) \mathrm{d}\hat{P}_D^*\left(D\right). \end{aligned} \tag{4}$$

  Given the assumptions on the distribution of the unobservable individual characteristics (ε and ν), we can compute the integral, either analytically or numerically.

2.2. Distribution assumptions

- Logit: setting $\theta_2 = 0$ and i.i.d. extreme values for ε_{ijt}.
- Nested Logit: setting $\theta_2 = 0$ and a nested structure for ε_{ijt}.
- Logit and Nested Logit fails to describe some observed correlation among ε_{ijt} (inflexible own-price elasticities and cross-price elasticities). However, if we allow a flexible enough variance-covariance matrix for ε_{ijt}, there are too many parameters that need to be estimated.
- Mixed Logit use random coefficients and the term μ_{ijt} to depict correlation without introducing too much parameters. But it also introduced two challenges:
  - There is no longer a closed-form expression for the probability, which requires some simulation methods.

- We need to know information about the consumers' heterogeneity to simulate the probability (and thus the share).

3. Estimation

3.1. The data

- Data required to consistently estimate the model:
  - Market shares s_{jt} and prices p_{jt} in each market;
  - Product characteristics x_{jt};
  - Distribution of demographics $\hat{P}_D^*$.
- Market shares are defined using a quantity variable, which depends on the context and should be determined by the specifics of the problem. Probably the most important consideration in choosing the quantity variable is the need to define a market share for the outside good. This share will rarely be observed directly, and will usually be defined as the total size of the market minus the shares of the inside goods. The total size of the market is assumed according to the context. When looking at historical data one can use eventual growth to learn about the potential market size. For example, one can assume that the market size is proportional to the size of the population with the proportionality factor equal to a constant factor, which can be estimated (Berry et al., 1996).
- Product characteristics include physical product characteristics and market segmentation information. They can be collected from manufacturer's descriptions of the product, the trade press, or the researcher's prior.
- The last component of the data is information regarding the demographics of the consumers in different markets. Unlike market shares, prices, or product characteristics, this estimation can proceed without demographic information. In this case the estimation will rely on assumed distributional assumptions rather than empirical distributions. But if we can include information about the distribution of demographics, it reduces the reliance on parametric assumptions. Therefore, instead of letting a key element of the method, the distribution of the random coefficients, be determined by potentially arbitrary distributional assumptions, we bring in additional information.

3.2. Identification

- An ideal experiment for identifying parameters of attributes—such as price effects—would randomly assigned different prices to consumers and record their purchasing patterns. This approach leverages exogenous price variation across markets for identification.
- Parameters of attributes in discrete choice models can be estimated using data from just one market, thus we are not mimicking the experiment. Instead, discrete choice models identifies parameters based on consumers' choices among products, which perceived as bundles of attributes. Therefore, the data from each market should not be seen as one observation of purchases when faced with a particular price vector; rather, it is an observation of the relative likelihood of purchasing J different bundles of attributes. This source of variation introduces some relevant issues:
  - Prices are not randomly assigned; rather, they are set by profit-maximization firms that take into account information that the research has to include in the error term (i.e., ξ_{jt}, since other unobserved components are integrated out in Equation (4)). Similar concerns apply to the identification of the effects of other attributes. This problem can be solved by using instrumental variables.
  - If one wants to tie demographic variables (distributions) to observed purchases, several markets, with variation in the distribution of demographics, have to be observed.

- Unlike parameters of attributes, identify the parameters that govern the distribution of the random coefficients relies on data from more markets. Furthermore, observing the change in market shares as new products enter or as characteristics of existing products change provides variation that is helpful in the estimation.

3.3. The estimation algorithm: Outline

- A straightforward approach to the estimation is to solve

$$\min_{\theta} \|s\left(x, p, \delta\left(x, p, \xi; \theta_1\right); \theta_2\right) - S\|, \tag{5}$$

  where $s(\cdot)$ are the market shares given by Equation (4), and S is the observed market shares. However, this approach is usually not taken, because all the parameters enter the minimization in Equation (5) in a nonlinear fashion and it is inconvenient to handle endogeneity.
- Let $Z = [z_1, \cdots, z_M]$ be a set of instruments such that

$$\mathbb{E}\left[z_m \omega\left(\theta^*\right)\right] = 0, \quad m = 1, \cdots, M, \tag{6}$$

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

$$\hat{\theta} = \arg\min_{\theta} \omega(\theta)' Z \Phi^{-1} Z' \omega\left(\theta\right), \tag{7}$$

  where Φ is a consistent estimate of $\mathbb{E}[Z'\omega\omega'Z]$.
- The error term is not defined as the difference between the observed and predicted market shares; rather, it is defined as the structure error ξ_{jt}. Note that ξ_{jt} only enters the mean utility level, $\delta(\cdot)$. Furthermore, the mean utility level is a linear function of ξ_{jt}; thus, in order to obtain an expression for the error term, we need to express the mean utility as a linear function of the variables and parameters of the model. In order to do this, we solve for each market the implicit system of equations

$$s\left(\delta_{\cdot t}; \theta_2\right) = S_{\cdot t}, \quad t = 1, \cdots, T, \tag{8}$$

  where $s(\cdot)$ are the market shares given by Equation (4), and S are the observed market shares. In solving this system of equations we have two steps:
  1. We need a way to compute the left-hand side of equation (8), which is defined by equation (4). For some special cases of the general model (e.g., logit, nested logit, and PD GEV) the market-share equation has a close-form expression. For the full random-coefficients model the integral defining the market shares has to be computed by simulation. The most common is to approximate the integral given by equation (4) by

$$\begin{aligned} s_{jt}(p_{\cdot t}, x_{\cdot t}, \delta_{\cdot t}, P_{\mathrm{ns}}; \theta_2) &= \frac{1}{\mathrm{ns}} \sum_{i=1}^{\mathrm{ns}} s_{jti} \\ &= \frac{1}{\mathrm{ns}} \sum_{i=1}^{\mathrm{ns}} \frac{\exp\left[\delta_{jt} + \sum_{k=1}^{K} x_{jt}^k \left(\sigma_k \nu_i^k + \pi_{k1} D_{i1} + \cdots + \pi_{kd} D_{id}\right)\right]}{1 + \sum_{m=1}^{J} \exp\left[\delta_{mt} + \sum_{k=1}^{K} x_{mt}^k \left(\sigma_k \nu_i^k + \pi_{k1} D_{i1} + \cdots + \pi_{kd} D_{id}\right)\right]}, \end{aligned} \tag{9}$$

     where $(\nu_i^1, \cdots, \nu_i^K)$ and $(D_{i1}, \cdots, D_{ik})$, $i = 1, \cdots, \mathrm{ns}$, are drawn from $\hat{P}_{\nu}^*$ and $P_D^*(D)$, respectively, while x_{jt}^k, $k = 1, \cdots, K$ are the variables that have random slope coefficients. Note that we use the extreme-value distribution $P_{\varepsilon}^*(\varepsilon)$, to integrate the ε's analytically.
  2. Note that, with these simulated integrals, δ_{jt} are the only unknown terms in Equation (8), thus we can invert the system of equations. For the full random-coefficients model, the system of Equations

(8) is nonlinear and is solve numerically. It can be solved by using the contraction mapping suggested by BLP, which amounts to computing the series

$$\delta_{\cdot t}^{h+1} = \delta_{\cdot t}^{h} + \ln S_{\cdot t} - \ln s\left(p_{\cdot t}, x_{\cdot t}, \delta_{\cdot t}^{h}, P_{\text{ns}}; \theta_2\right), \quad t = 1, \cdots, T, \quad h = 1, \cdots, H-1, \qquad (10)$$

where H is the smallest integer such that $\|\delta_{\cdot t}^{H} - \delta_{\cdot t}^{H-1}\|$ is smaller than some tolerant level, and $\delta_{\cdot t}^{H}$ is the approximation to $\delta_{\cdot t}$.

- Once the inversion has been computed, either analytically or numerically, the error term is defined as

$$\omega_{jt} = \delta_{jt}\left(S_{\cdot t}; \theta_2\right) - \left(x_{jt}\beta + \alpha p_{jt}\right) \equiv \xi_{jt}. \qquad (11)$$

  We then search for $\hat{\theta}$ that minimizes the sample analog of moment conditions in Equation (6).
- Taken together, there are essentially five steps to follow in computing the estimates:
  1. Prepare the data including draws from the distribution of ν and D;
  2. For a given value of θ_2 and δ, compute the market shares implied by Equation (9);
  3. For a given value of θ_2, compute the vector δ that equates the market shares computed in Step 2 to the observed shares using Equation (10).
  4. For a given θ, compute the error term using Equation (11) (as a function of the mean valuation computed in Step 3), interact it with the instruments, and compute the value of the objective function in Equation (7).
  5. Search for the value of θ that minimizes the objective function computed in Step 4.

3.4. The estimation algorithm: Details

- *Step 1:* Prepare the data including draws from the distribution of ν and D.
  - It is useful to define a vector of market shares and two matrices of "right-hand size" variables, X_1 and X_2. The first, X_1, contains the variables that enter $\delta(\cdot)$, which is common to all individuals within the market. The latter, X_2, contains the variables that will have a random coefficient. X_1 and X_2 could be different.
  - We then sample a set of "individuals." Each "individual" consists of a K-dimensional vector of shocks that determine the individual's taste parameters, $\nu_i = (\nu_i^1, \cdots, \nu_i^K)$, demographics, $D_i = (D_{i1}, \cdots, D_{id})$, and potentially a vector of shocks to the utility, $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iJ})$. Let ns be the number of individuals sampled. In most cases we will not draw ε_i since we can integral the extreme value shocks analytically.
  - Shocks that determine the individuals taste parameters, ν_i, are usually drawn from a multi-variate normal distribution. In principle, these could be drawn from any other parametric distribution. The choice of the distribution depends on the problem at hand and on the researcher's beliefs.
  - Demographics, D_i, are drawn from survey or census data; instead of assuming parametric forms for the distribution of demographics, real individuals are sampled.
  - Three questions relevant to the sampling:
    - In theory we could consider drawing different individuals for each observation, i.e., each brand in each market. However, in order for Step 3 to work we require that the predicted market shares sum up to one, which requires that we use the same draws for each market.
    - Whether these draws should vary between markets depends on the specifics of the problem and the data. In general, the draws should be the same whenever it is the same individuals making the decisions. For example, BLP use national market shares of different brands of cars over twenty years. They use the same draws for all markets.
    - It is important to draw these only once at the beginning of the computation. If the draws are changed during the computation the non-linear search is unlikely to converge.

- *Step 2:* For a given value of θ_2 and δ, compute the market shares implied by Equation (9).
  - Now that we have a sample of individuals, each described by $(\nu_i, D_i, \varepsilon_i)$, for a given value of the parameters, θ_2, and the component common to all consumers, δ, we compute the predicted market shares given by the integral in Equation (4). For many models (e.g., Logit, Nested Logit and PD GEV), this step can be performed analytically.
  - For the full random-coefficients model, this integral can be computed by simulation. It is very recommended to review the simulation methods introduced in Chapter 5 of Train's book (a note is here).
    - The first approach is the naive frequency estimator (i.e., accept–reject simulator), given by

$$s_{jt}\left(p_{jt}, x_{jt}, \delta_{\cdot t}, P_{ns}; \theta_2\right) = \frac{1}{\mathrm{ns}} \sum_{i=1}^{\mathrm{ns}} \mathbb{1}\left\{u_{ijt}\left(D_i, \nu_i, \varepsilon_i\right) \geq u_{ilt}\left(D_i, \nu_i, \varepsilon_i\right),\ \forall l = 0, 1, \cdots, J\right\}.$$

      This is repeated in every market. However, as mentioned in Train (2009), there are two drawbacks of this method. First, it requires a large number of draws, ns, to ensure that all predicted market shares are not zero. Second, the objective function is not smooth, which does not allow us to use gradient methods to minimize the objective function (see Step 5).
    - The second approach is the smooth simulator. We use the logit function to smooth the indicator function, and the predicted market shares are approximated by

$$s_{jt} = \frac{1}{\mathrm{ns}} \sum_{i=1}^{\mathrm{ns}} \frac{\exp\left[\delta_{jt} + \sum_{k=1}^{K} x_{jt}^{k}\left(\sigma_k \nu_i^k + \pi_{k1} D_{i1} + \cdots + \pi_{kd} D_{id}\right)\right]}{1 + \sum_{m=1}^{J} \exp\left[\delta_{mt} + \sum_{k=1}^{K} x_{mt}^{k}\left(\sigma_k \nu_i^k + \pi_{k1} D_{i1} + \cdots + \pi_{kd} D_{id}\right)\right]},$$

      which is exactly Equation (9). This approach is equivalent to assuming that ε_i are drawn from the type-I extreme value distribution. As the conditional probabilities can be calculated analytically, we do not need to draw ε_i in Step 1.
- *Step 3:* For a given value of θ_2, compute the vector δ that equates the market shares computed in Step 2 to the observed shares.
  - We want to compute the $J \times T$-dimensional vector of mean valuations, δ_{jt}, that equates the market shares computed in Step 2 to the observed shares. This amounts to solving separately for each market the system of equations:

$$s\left(\delta_{\cdot t}; \theta_2\right) = S_{\cdot t}, \quad t = 1, \cdots, T. \tag{12}$$

    For the logit model, this inversion can be computed analytically by $\delta_{jt} = \ln S_{jt} - \ln S_{0t}$, where S_{0t} is the market share of the outside good. Therefore, we can solve a logit model without simulation.
  - For the full model, the system of equations is non-linear and is solved numerically. It can be solved by using the contraction mapping suggested by BLP:

$$\delta_{\cdot t}^{h+1} = \delta_{\cdot t}^{h} + \ln S_{\cdot t} - \ln s\left(p_{\cdot t}, x_{\cdot t}, \delta_{\cdot t}^{h}, P_{\mathrm{ns}}; \theta_2\right), \quad t = 1, \cdots, T, \quad h = 1, \cdots, H-1,$$

    where H is the smallest integer such that $\|\delta_{\cdot t}^{H} - \delta_{\cdot t}^{H-1}\|$ is smaller than a pre-set tolerant level, and $\delta_{\cdot t}^{H}$ is the approximation to $\delta_{\cdot t}$.
  - Convergence can be reached faster by choosing a good starting value. It is recommended to use the values that solve for the logit model; that is,

$$\delta_{jt}^{0} = \ln S_{jt} - \ln S_{0t}.$$

- To reduces the number of exponents and logarithms computed, the contraction mapping can be solved by solving for the exponent of the vector δ. Define $w_{jt} = \exp(\delta_{jt})$, the iteration becomes

$$w_{jt}^{h+1} = w_{jt}^{h} \frac{S_{jt}}{s_{jt}\left(p_{\cdot t}, x_{\cdot t}, \delta_{\cdot t}^{h}, P_{\text{ns}}; \theta_2\right)}, \quad t = 1, \cdots, T, \quad h = 1, \cdots, H-1,$$

  where the starting value is $w_{\cdot t}^{0} = \exp(\delta_{\cdot t}^{0})$ and $w_{\cdot t}^{H} = \exp(\delta_{\cdot t}^{H})$ is the same as the solution to the initial contraction mapping.

- *Step 4:* For a given θ_2, compute the error term, interact it with the instruments, and compute the value of the objective function in Equation (7).
  - Compute the error term: $\omega = \delta - X_1\theta_1$. Notice that given θ_2, θ_1 can be analytically derived (see Step 5). The objective function is $\omega(\theta)' Z \Phi^{-1} Z' \omega(\theta)$, where Φ is a consistent estimate of $\mathbb{E}[Z'\omega\omega' Z]$. To derive it, we need knowledge of the weight matrix.
    - We can assume homoskedastic errors and therefore the optimal weight matrix is proportional to $Z'Z$.
    - We can compute an estimate of θ, say $\hat{\theta}$, using $\Phi = Z'Z$, and then use this estimate to compute a new weight matrix $\mathbb{E}[Z'\omega\omega' Z]$, which in turn is used to compute a new estimate of θ.

- *Step 5:* Search for the value of θ_2 that minimizes the objective function computed in Step 4.
  - For the logit model, this searching can be done analytically as it is a linear GMM problem. For the full model, we need to perform a non-linear search over θ. The first-order condition with respect to θ_1 is

$$X_1' Z \Phi^{-1} Z' \left(\delta - X_1 \hat{\theta}_1\right) = 0,$$

    which implies that

$$\hat{\theta}_1 = \left(X_1' Z \Phi^{-1} Z' X_1\right)^{-1} X_1' Z \Phi^{-1} Z' \delta, \tag{13}$$

    where $\delta = \delta(\hat{\theta}_2)$. Then the non-linear search can be limited to θ_2.
  - One of two search methods is usually used, either the Nelder–Mead (1965) non-derivative "simplex" search method, or a quasi-Newton method with an analytic gradient (see Press et al., 1994, and references therein). The first is more robust but is much slower to converge, while the latter is two orders of magnitude faster, yet is sensitive to starting values. The recommended practice is to start with the non-derivative method and switch to the gradient method once the objective has been lowered to reasonable levels.
  - To use the gradient method, we need to calculate the gradient of the objective function $\omega(\theta)' Z \Phi^{-1} Z' \omega(\theta)$ with respect to θ_2, which is

$$2\left(\frac{\partial \omega\left(\theta\right)}{\partial \theta_2}\right)' Z \Phi^{-1} Z' \omega\left(\theta\right), \tag{14}$$

    and

$$\frac{\partial \omega\left(\theta\right)}{\partial \theta_2} = \frac{\partial \delta}{\partial \theta_2} - \left(\frac{\partial \hat{\theta}_1}{\partial \theta_2}\right)' X_1'. \tag{15}$$

    Now substitute Equation (15) into (14) and find $X_1' Z \Phi^{-1} Z' \omega(\theta) = X_1' Z \Phi^{-1} Z' (\delta - X_1 \hat{\theta}_1) = 0$. Therefore, the gradient can be further simplified as

$$2\mathrm{D}\delta' Z \Phi^{-1} Z' \omega,$$

where $\mathrm{D}\delta \equiv \partial\delta/\partial\theta_2$. As observed in Equation (12), δ is defined by the system of J equations, thus we can use the Implicit Function Theorem to yield

$$\mathrm{D}\delta \equiv \begin{bmatrix} \frac{\partial \delta_{1t}}{\partial \theta_{21}} & \cdots & \frac{\partial \delta_{1t}}{\partial \theta_{2L}} \\ \vdots & & \vdots \\ \frac{\partial \delta_{Jt}}{\partial \theta_{21}} & \cdots & \frac{\partial \delta_{Jt}}{\partial \theta_{2L}} \end{bmatrix} = - \begin{bmatrix} \frac{\partial s_{1t}}{\partial \delta_{1t}} & \cdots & \frac{\partial s_{1t}}{\partial \delta_{Jt}} \\ \vdots & & \vdots \\ \frac{\partial s_{Jt}}{\partial \delta_{1t}} & \cdots & \frac{\partial s_{Jt}}{\partial \delta_{Jt}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial s_{1t}}{\partial \theta_{21}} & \cdots & \frac{\partial s_{1t}}{\partial \theta_{2L}} \\ \vdots & & \vdots \\ \frac{\partial s_{Jt}}{\partial \theta_{21}} & \cdots & \frac{\partial s_{Jt}}{\partial \theta_{2L}} \end{bmatrix},$$

where θ_{2i}, $i = 1, \cdots, L$ denotes the i's element of the vector θ_2. These derivatives are

$$\begin{aligned} \frac{\partial s_{jt}}{\partial \delta_{jt}} &= \frac{1}{\mathrm{ns}} \sum_{i=1}^{\mathrm{ns}} s_{jti}\left(1 - s_{jti}\right), \\ \frac{\partial s_{jt}}{\partial \delta_{mt}} &= -\frac{1}{\mathrm{ns}} \sum_{i=1}^{\mathrm{ns}} s_{jti} s_{mti}, \\ \frac{\partial s_{jt}}{\partial \sigma_k} &= \frac{1}{\mathrm{ns}} \sum_{i=1}^{\mathrm{ns}} \nu_i^k s_{jti} \left(x_{jt}^k - \sum_{m=1}^{J} x_{mt}^k s_{mti} \right), \\ \frac{\partial s_{jt}}{\partial \pi_{kd}} &= \frac{1}{\mathrm{ns}} \sum_{i=1}^{\mathrm{ns}} D_{id} s_{jti} \left(x_{jt}^k - \sum_{m=1}^{J} x_{mt}^k s_{mti} \right). \end{aligned}$$

3.5. Instruments

- A standard place to start the search for demand-side instrumental variables is to look for variables that shift cost and are uncorrelated with the demand shock. These are the textbook instrumental variables, which work quite well when estimating demand for homogeneous products. The problem with the approach is that we rarely observe cost data fine enough that the cost shifters will vary by brand.
- The most popular identifying assumption used to deal with the above endogeneity problem is to assume that the location of products in the characteristics space is exogenous, or at least determined prior to the revelation of the consumers' valuation of the unobserved product characteristics. Built on similar assumption, BLP use the observed product characteristics (excluding price and other potentially endogenous variables), the sums of the values of the same characteristics of other products offered by that firm (if the firm produces more than one product), and the sums of the values of the same characteristics of products offered by other firms.
- The last set of instrumental variables exploit the panel structure of the data. The identifying assumption made is that, controlling for brand-specific intercepts and demographics, the city-specific valuations of the product, $\Delta\xi_{jt} = \xi_{jt} - \xi_j$, are independent across cities but are allowed to be correlated within a city over time. Given this assumption, the prices of the brand in other cities are valid instruments.
- The extent to which the assumptions needed to support any of the above instrumental variables are valid in any given situation is an empirical issue. Resolving this issue beyond any reasonable doubt is difficult and requires comparing results from several sets of instrumental variables, combing additional data sources, and using the researcher's knowledge of the industry.

3.6. Brand-specific dummy variables

- If enough markets are observed, then brand-specific dummy variables ξ_j can be included as product characteristics. They should be used whenever possible.

- With brand-specific dummies, the error term is the market-specific deviation from the mean valuation $\Delta\xi_{jt} = \xi_{jt} - \xi_j$. The inclusion of brand dummy variables introduces a challenge in estimating the taste parameters β.
- A relevant concern is that if product characteristics are fixed across market, the taste parameters may not be identified. However, we can retrieve these parameters with some assumptions. Suppose that $d = (d_1, \cdots, d_J)'$ is the $J \times 1$ vector of brand dummies, X is the $J \times K$-dimensional matrix of characteristics that are fixed across market, and $\xi = (\xi_1, \cdots, \xi_J)$ are effects of unobserved attributes captured by brand-specific fixed effects. Accordingly, we have

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

  If we assume that $\mathbb{E}[\xi \mid X] = 0$, then the estimates of β and ξ are

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

  where $\hat{d}$ is the vector of coefficients estimated and V_d is the variance-covariance matrix of these estimates.