Discrete Choice Models

Basics, logit, and BLP

Ashvin Gandhi[1]

Harvard University

September 22, 2015[2]

[1] agandhi@fas.harvard.edu

[2] Based on previous notes by Daniel Pollmann, Tom Wollmann, and Michael Sinkinson.

Agenda

- Review demand systems
- Explore approaches to overcoming IIA: Nested Logit, Multinomial Probit, and BLP
- BLP: Details in estimation, and implications
- Compare and contrast with Pure Characteristics model

Demand Systems: Story So Far

- Characteristic space: products are bundles of characteristics.
- Consumers choose the bundle that maximizes their utility.
- Hotelling: product characteristic is location, I want the "closest" product to me, to minimize transport costs. Canonical models: linear city, circular city, Median Voter Theorem.
- Vertical: we all agree on the rankings, but products differ in price and we all dislike paying at varying degrees.

  $$U_{ij} = \delta_j - v_i p_j$$

Vertical model

$$U_{ij} = \delta_j - v_i p_j$$

- Substitution patterns? Only with "neighbor" products.
- Estimation? Recover δ_j values from prices, shares and an assumption on v.
  - If we know the distribution of v, we can "invert" market shares for the ranges of v that purchase each good by lining consumers up in order of taste for quality/MU; cut-offs imply quality index δ.
- Application: Bresnahan (JIE, 1987)

Aside: "Almost Ideal Demand System"

- AIDS: "Almost Ideal Demand System" (Deaton and Muellbauer, AER, 1980)
- Consumers divide budget over different categories of goods.
- Derived from Piglog model, which allows researchers to treat aggregate consumer behavior as if it were the outcome of a single maximizing consumer.
- Appeal (at the time): satisfies the axioms of choice, aggregates over consumers without invoking parallel linear Engel curves, consistent with budget constraints, simple to estimate.
- Example: Hausman, Leonard and Zona (Annales d'Économie et de Statistique, 1994)

Pure Logit

- Each product has a "mean" quality level, and we all have private tastes:

$$U_{ij} = \tilde{\delta}_j - \alpha p_j + \epsilon_{ij}$$

- Term ϵ_{ij} is Type-I Extreme Value (Gumbel) distribution ("logit errors"), $F(\epsilon) = \exp\left[-\exp\left(-\epsilon\right)\right]$.
  - **Why do we use it?**[3]

[3] See also Jaibi and ten Raa (1998) and Matejka and McKay (AER, 2015) for micro foundations.

Pure Logit shares and derivatives

- Normalize U of outside good to 0. Let $\delta_j = \tilde{\delta}_j - \alpha p_j$. Shares:

$$
\begin{aligned}
s_j(\theta) &= \frac{\exp[\delta_j]}{1 + \sum_q \exp[\delta_q]} \\
s_0(\theta) &= \frac{1}{1 + \sum_q \exp[\delta_q]}
\end{aligned}
$$

- Analytic form for price derivatives:

$$
\begin{aligned}
\frac{\partial s_j}{\partial p_j} &= -\alpha s_j (1 - s_j) \\
\frac{\partial s_j}{\partial p_k} &= \alpha s_j s_k
\end{aligned}
$$

  - Definition of outside good/potential market matters

Logit estimation

- Recover mean utility (Berry, RAND, 1994):

$$\delta_j = \ln\left[s_j(\theta)\right] - \ln\left[s_0(\theta)\right]$$

- We observe characteristics k of product j, so we can model it as:

$$\delta_j = \sum_k x_{kj}\beta_k - \alpha p_j + \xi_j$$

- What is ξ_j?
  - Unobserved heterogeneity (to the econometrician); here, structural error: unobserved product characteristic
  - Mean-independent of characteristics, aside from price (since firms know ξ_j when setting price)
- Why is ξ_j important to include?
  - Allows us to rationalize any pattern of market shares

IIA problem

- The latent utility in the logit model is:

  $$U_{ij} = X_{ij}'\beta + \epsilon_{ij}$$

  We assume that the idiosyncratic taste term, ϵ_{ij}, is independent across all individuals and choices. However, it is reasonable to assume that these tastes may be correlated, especially between similar products.
- McFadden's (1981) example: Commuting choices are "car", "red bus", and "blue bus". We would expect tastes for red bus and blue bus to be positively correlated. However, in the Pure Logit model, if you remove the "blue bus" choice, those that chose the blue bus would now split between "car" and "red bus" according to their relative shares.

IIA problem: Some Alternatives

- Some possible solutions:
  - Researcher can select groups of products that we would expect to have correlated tastes for (e.g., Nested Logit)
  - Researcher can attempt to estimate correlation between taste terms (e.g., Multinomial Probit)

Nested Logit (1/2)

- We are going to divide the set of possible choices into "nests" (which may themselves contain sub-nests).
- The key difference in this model is that we will now assume that the ϵ_{ij} for all choices within a nest have a correlation coefficient of $\left(1-\rho^2\right)$, but that the correlation between nests is 0. The researcher will place choices that are believed to be correlated in the same nest.
- Estimating these types of models can be done via Maximum Likelihood, or "backwards induction": Within each nest, we are back to a standard Conditional Logit model, except that our estimated parameters are transformed (β/ρ). Solve each nest, then step up a level and estimate the choice between nests.
- McFadden (1981) provides details on two-step estimator and inference.

Nested Logit (2/2)

- The researcher must "pick" the nests. This is unattractive: we would prefer the data to "tell" us the nests.
- The nests matter: different nest structures may produce significantly different results.
- We estimate ρ, which tells us the estimated correlation between unobserved components within nests.
- Most useful for modeling choices that have a sequential nature (e.g., Goldberg, EMA, 1995).
    - For example, a person may first decide between a laptop computer and a desktop computer. Then, within laptop, they may choose between an "ultraportable" and a "desktop replacement" laptop. Within each of those nests are 10-15 different options.
    - In that case, substitution patters will be stronger for products within the same nest than between products in different nests.
    - How would this work for cars?

Multinomial Probit

- The most straightforward way to allow correlation between choices is to actually free up the covariance matrix of the unobserved term between choices. Formally, we now let

  $$\epsilon_i | X_i \sim N(\mu, \Omega)$$

- The covariance matrix Ω is a (symmetric) matrix of size $(J+1)$. Note that we are using a Normal distribution, in contrast to the previous models.
- From a practical perspective, μ usually restricted to 0 for all choices.
- There are typically other restrictions required in order to estimate the covariance matrix with any precision. This approach adds a large number of parameters to be estimated, especially as J grows large (the symmetric covariance matrix with J goods and an outside good has $\frac{(J+1)(J+2)}{2}$ parameters).

Multinomial Probit considerations

- Historically very hard to compute as J grows large (requires precise evaluation of a $J+1$-dimensional MVN distribution).
- Numerical methods and technology have advanced to make this more feasible (e.g, GHK simulator), but data requirements still substantial.
- Stata commands: "mprobit" (ML), "asmprobit" (GHK). Can specify restrictions on covariance matrix.
- The Nested Logit case is like a version of the MNP where the errors are instead distributed type-I extreme value, and the covariance matrix is restricted so that goods within the same nest have correlation coefficient $\left(1-\rho^2\right)$ and goods across nests have 0.
- Reference on simulation methods for discrete choice: Train (2nd ed., 2009), also available for download from Kenneth Train's website.

Multinomial Probit example

- Useful when arbitrary correlation is an important feature of the data and central to the question of interest.
- Example: Chu, Leslie and Sorensen (AER, 2011). Estimating demand for theater tickets at a local theater company ($J = 220$, all possible bundles to 8 different plays, including some "subscriptions"). Estimate this type of model using method of moments due to data issues.
- Paper was examining bundle pricing, a question where correlation between tastes strongly impacts appeal of bundles.

BLP introduction

BLP can be estimated using macro data (market shares), micro data, or a combination of both. We observe for $j = 1, \dots, J$ products:

- Product characteristics X_{jk}, $k = 1, \dots, K$
- Price p_j
- Market share s_j^o (also s_0^o)
- Possibly other “moments”
- What we assume:
  - ξ_j, ϵ_{ij} as in Logit
  - Often a pricing equation (Nash)

BLP utility: Random Utility Logit (1/4)

$$
\begin{aligned}
U_{ij} &= \sum_k x_{jk}\beta_{ik} + \xi_j + \epsilon_{ij} \\
\beta_{ik} &= \lambda_k + \beta_k^{o\prime} z_i + \beta_k^{u\prime} v_i
\end{aligned}
$$

Alternative notation is:

$$
\begin{aligned}
U_{ij} &= X_j\beta_i + \xi_j + \epsilon_{ij} \\
\beta_i &= \lambda + \Sigma_z z_i + \Sigma_\nu v_i
\end{aligned}
$$

where the k-th row ofΣ_z and $\Sigma_r$are $\beta_k^{o\prime}$and $\beta_k^{u\prime}$, respectively.

BLP utility: Random Utility Logit (2/4)

$$
\begin{aligned}
U_{ij} &= \sum_k x_{jk}\beta_{ik} + \xi_j + \epsilon_{ij} \\
\beta_{ik} &= \lambda_k + \beta_k^{o\prime} z_i + \beta_k^{u\prime} v_i
\end{aligned}
$$

- Interaction terms: consumers with certain characteristics prefer similar product characteristics (addresses IIA)
  - $\beta_k^{o\prime}$: parameterize household characteristics (z_i). Note: frequently, mico-level choice data are not available, so often fit to census distribution. If micro-level choice data available (e.g. "micro"-BLP, 2004), one can leverage choice covariances with household characteristics.
  - $\beta_k^{u\prime}$: parametrize vector of random draws (v_i), which are usually standard normal.

BLP utility: Random Utility Logit (3/4)

$$\begin{aligned} U_{ij} &= \sum_k x_{jk}\beta_{ik} + \xi_j + \epsilon_{ij} \\ \beta_{ik} &= \lambda_k + \beta_k^{o\prime} z_i + \beta_k^{u\prime} v_i \end{aligned}$$

- Unobserved product quality (ξ_j): known to consumers and firms but not econometrician (structural error)
- Type-I EV iid error term: analytic shares and derivatives

BLP utility: Random Utility Logit (4/4)

$$
\begin{aligned}
U_{ij} &= \sum_k x_{jk}\beta_{ik} + \xi_j + \epsilon_{ij} \\
\beta_{ik} &= \lambda_k + \beta_k^{o\prime} z_i + \beta_k^{u\prime} v_i
\end{aligned}
$$

- Can also be seen in the "quality" framework using δ:

$$
\begin{aligned}
U_{ij} &= \delta_j + \sum_{k,r} x_{j,k} z_{i,r} \beta^o_{k,r} + \sum_{k,l} x_{j,k} v_{i,l} \beta^u_{k,l} + \epsilon_{ij} \\
\delta_j &= \sum_k x_{j,k}\lambda_k + \xi_j
\end{aligned}
$$

BLP features

- How do these added interaction terms address the IIA problem?
  - In Nested Logit and MNP, we found ways to have some products "closer" to each other by modifying the ϵ_{ij} term.
  - Now, products are "close" based on the types of consumers that purchase them; if a given product were removed from the choice set, consumers would substitute to other goods that are close in terms of characteristics.
  - At the individual level, IIA still holds, but we have heterogeneous consumers, so IIA is broken at the aggregate level.
- What is the price to pay?
  - The shares are no longer easy to calculate as in the Logit model. However, BLP provides an estimation algorithm involving simulation and a contraction mapping.

BLP shares

Let's look at a simple version with only unobserved interactions.

- Normalize U of outside good to 0. Shares:

$$
\begin{aligned}
s_j(\theta) &= \int \underbrace{\frac{\exp\left[\delta_j + \sum_{k,l} x_{j,k} v_l \beta^u_{k,l}\right]}{1 + \sum_q \exp\left[\delta_q + \sum_{k,l} x_{q,k} v_l \beta^u_{k,l}\right]}}_{\equiv s_j(\theta;\nu)} dF(\nu) \\
s_0(\theta) &= \int \frac{1}{1 + \sum_q \exp\left[\delta_q + \sum_{k,l} x_{q,k} v_l \beta^u_{k,l}\right]} dF(\nu)
\end{aligned}
$$

BLP derivatives

Simple version with only unobserved interactions.

- Analytic form for price derivatives:

$$\begin{aligned} \frac{\partial s_j}{\partial p_j} &= -\int (\alpha + \sigma_\alpha \nu_\alpha) \, s_j(\theta; \nu) \left[1 - s_j(\theta; \nu)\right] dF(\nu) \\ \frac{\partial s_j}{\partial p_k} &= \int (\alpha + \sigma_\alpha \nu_\alpha) \, s_j(\theta; \nu) \, s_k(\theta; \nu) \, dF(\nu) \end{aligned}$$

  - Cannot just multiply average shares!
  - Definition of outside good/potential market matters

BLP program: Objective

- To estimate $\theta = (\beta^u, \lambda)$, solve the following problem for some weighting matrix W:

$$\min_{\theta,\xi} g(\xi)' W g(\xi)$$
$$\text{s.t. } \tilde{s}(\theta, \xi) = s^o, \tag{1}$$

  for some moment conditions $\mathbb{E}[g(\xi)] = 0$. The question is what $g(\xi)$should be and how it can be computed.

- The equilibrium constraints (1) imply, for each value of θ, a vector of mean utilities $\delta(\theta)$ such that with $\xi_j = \delta_j(\theta) - \sum_k x_{j,k}\lambda_k$, $\tilde{s}(\theta, \xi) = s^o$ is satisfied (predicted shares equal to actual shares). This vector $\delta(\theta)$ is unique for a given β and can be found using the BLP contraction mapping.

BLP program: Objective

- This uniqueness allows us to rewrite the objective as:

$$\min_{\theta} \underbrace{g\left(\xi\left(\theta\right)\right)' W g\left(\xi\left(\theta\right)\right)}_{Q_W(\theta)},$$

where $g\left(\xi\right)$ as before, and $\xi_j\left(\theta\right) = \delta_j\left(\beta\right) - \sum_k x_{j,k}\lambda_k$ with $\delta\left(\beta\right)$ such that $\tilde{s}\left(\beta, \delta(\beta)\right) = s^o$. The mean utility vector $\delta\left(\beta\right)$ is found using the contraction mapping in BLP.

BLP program: Inversion

- Iterate contraction mapping to get you $\delta(\beta)$ from the simulated shares (tol. $\leq 10^{-8}$).

$$\begin{aligned} \delta_j^N(\beta) &= \delta_j^{N-1}(\beta) + \ln\left[s_j^o\right] - \ln\left[\hat{s}_j^{ns}\left(\beta, \delta^{N-1}\right)\right] \\ \|\delta_j^N(\beta) - \delta_j^{N-1}(\beta)\| &\to 0 \end{aligned}$$

- Tip: Choosing a good $\delta_j^0(\beta)$can greatly speed up your code.

BLP program: Simulation

- To simulate shares conditional on (β, δ): take *ns* standard Normal draws for v, Type-I EVD gives analytic shares for each draw, sum over draws:

$$\hat{s}_j^{ns}(\theta, \delta) = \frac{1}{ns} \sum_{i=1}^{ns} \frac{\exp\left[\delta_j + \sum_k x_{j,k} v_{i,k} \beta_k^u\right]}{1 + \sum_q \exp\left[\delta_q + \sum_k x_{q,k} v_{i,k} \beta_k^u\right]}$$

- Why is this better than the naive simulation method? (Hint: Smoothness.)
- Important tip: use the same draws of v. Why? (Hint: Smoothness.)

BLP program: Approach

- Computation can be substantially simplified by separating $\theta = (\theta_1, \theta_2)$, here, $\theta_1 = \lambda$ and $\theta_2 = \beta$. Search only over nonlinear parameters θ_2; for each θ_2, θ_1 is given ("concentrated out") by GMM/2SLS of $\xi_j\left(\lambda, \delta_j(\beta)\right) = \delta_j\left(\beta\right) - \sum_k x_{jk}\lambda_k$ on x with appropriate instruments.
- The BLP program is constructed to have in an inner and outer loop.
- How does 2 step GMM fit in?

BLP program: Outer Loop

- Outer loop: Minimize $Q_W\left(\lambda^*(\beta), \beta\right)$ over β .
  - We "concetrate out" $\theta_1 = \lambda^*(\beta) = argmin_\lambda Q_W(\lambda, \beta)$, we need only search over $\theta_2 = \beta$ when using fminsearch/fmincon/etc. So you can think of the problem as $\min_{\theta_2}\left[\min_{\theta_1} g\left(\xi\left((\theta_1, \theta_2)\right)\right)' Wg\left(\xi\left((\theta_1, \theta_2)\right)\right)\right]$, where the inner minimization problem has a simple (linear GMM) analytical solution given $\delta(\theta_2) \equiv \delta(\beta)$.
  - In order to concentrate out θ_1, we must know $\delta\left(\beta\right)$. This is the inner loop.
  - Must provide an intial value in the search for β^*. Quality of guess greatly affects how quickly the code converges.

BLP program: Inner Loop

- Inner loop: Get $\delta\left(\beta\right)$.
  - Iterate the contraction mapping $\delta_j^k\left(\theta\right) = \delta_j^{k-1}\left(\theta\right) + \ln\left[s_j^o\right] - \ln\left[\hat{s}_j^{ns}\left(\theta, \delta^{k-1}\right)\right]$ or its exponentiated version, where $\hat{s}_j^{ns}\left(\theta, \delta\right) = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\exp\left[\delta_j + \sum_k x_{j,k} v_{i,k} \beta_k^u\right]}{1 + \sum_q \exp\left[\delta_q + \sum_k x_{q,k} v_{i,k} \beta_k^u\right]}$.
  - You will need to provide an initial value for the contraction mapping. The stopping rule for the contraction mapping can be chosen as $|\max_j\left(\delta_j^k\left(\theta\right) - \delta_j^{k-1}\left(\theta\right)\right)| \leq \epsilon_{in}$ with $\epsilon_{in} = 10^{-8}$ or better even $\epsilon_{in} = 10^{-14}$.
  - Use the same draws of ν at every candidate value θ (Pakes and Pollard, EMA, 1989, p. 1048).

BLP Supply Side

- Usual firm FOC: $\frac{\partial s}{\partial p}(p - mc) + s(p) = 0$
- Multi-product firm FOC:
  $\Delta(p - mc) + s(p) = 0 \Leftrightarrow mc = \Delta^{-1} s(p) + p$
  - $\Delta_{jk} = \begin{cases} \frac{\partial s_k}{\partial p_j} & \text{if } f_j = f_k \\ 0 & \text{otherwise} \end{cases}$
- Parametrize $mc = x'\gamma + \omega$; then, $\gamma \in \theta_1$ (but $\alpha \in \theta_2$!), interact $\omega(\theta)$ with appropriate instruments, and add to GMM system.

BLP instruments

- Characteristics: for product j of firm f (for an arbitrary ordering of products within firms),
$$\left\{ z_{jk}, \sum_{r \neq j,\, r \in \mathcal{F}_f} z_{rk}, \sum_{r \neq j,\, r \notin \mathcal{F}_f} z_{rk} \right\},$$
where z_{jk} is kth characteristic of product j, $\mathcal{F}_f$ set of products by firm f (always constructed at market level)
    - Measure of isolation in product space
    - Approximation to optimal instruments
- Cost shifters
- Nevo (EMA, 2001) instruments: prices of the same brand in different markets
    - Valid with independent demand shocks, correlated cost shocks
    - Hausman instruments

Standard errors and optimal weighting matrix

- In GMM, we form a moment condition of the form $\mathbb{E}\left[\psi\left(w_j,\theta\right)\right]=0$.[4]
- Inference:
$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,V\right),$$
where $V=\left(\Gamma' C\Gamma\right)^{-1}\Gamma' C\Delta C\Gamma\left(\Gamma' C\Gamma\right)^{-1}$.
  - Estimate $\hat{\Gamma}$ using analytic derivative or finite differences.
  - Estimate of $\hat{\Delta}$ needs to account for simulation error in BLP-type problem.
- Optimal weighting matrix: $C\propto\Delta^{-1}$
  - Use estimate $\hat{\Delta}$ (next).

[4] Above, I used the notation $g\left(\cdot\right)$, because the argument was ξ. Some people use $g\left(\cdot\right)$ (e.g., Newey in 14.385 notes) instead of $\psi\left(\cdot\right)$ (e.g., Chamberlain in 2120 notes) – just notation, but be careful.

Covariance in BLP-type problems

- There are three sources of error in BLP:
  - V_1: "arises from the process generating the product characteristics (x, w, ξ, ω)."
  - V_2: "from the consumer sampling process."
  - V_3: "from the simulation process."
- See BLP and BLintonP for details.

Conclusion

- Problem set 1 assigned.
- Feel free to email me with questions. I'll do my best to get back to you promptly.