This lecture will cover four topics:

1. Application of the BLP framework with micro-level choice data (BLP, 2004) and A. Langer's model of gender-based price discrimination in the retail auto sector.
2. Aside: sorting effects in hedonic models
3. Neighborhood choice problems - Bayer Ferreira and McMillan
4. Nested logit and other extensions of MNL

1. Applications of the BLP two-stage procedure with micro-level data

As noted in the previous lecture, the BLP approach can be adapted to situations where the researcher has access to a micro sample. In the 2004 "microBLP" paper, and in Langer's setting, the micro sample is based on a survey of recent car buyers: buyers are asked which model they purchased, how much they paid for the car, the model that was their "second" choice, and a variety of demographic information. A difficulty is that the sample excludes information on people who did not buy cars. We will present a likelihood-based approach to addressing this problem (following Langer) after we decribe the easier case of seeing all choices.

Assume that the indirect utility that agent i obtains from buying product j is

$$u_{ij} = x_j(\overline{\beta} + \beta^0 z_i + \beta^\nu \nu_i) + \xi_j + \epsilon_{ij},$$

where x_j is a set of observed characteristics of product j (including price), z_i is a vector of observed characteristics of agent i (normalized to have mean 0), ν_i is a random unobserved taste shifter (possibly vector-valued) for agent i (assumed to follow some standard distribution, like a normal), ξ_j represents an unobserved attribute of product j (note that this has the same effect on all $i's$ – an assumption that might be problematic in some settings) and ϵ_{ij} is our old friend, the EV1 error. As usual in logit settings, assume that the indirect utility of not buying is

$$u_{i0} = \epsilon_{i0},$$

i.e., that $x_0 = \xi_j = 0$. In some cases this is highly restrictive. In the new car purchase case, for example, each consumer presumably has a different "default" associated with not buying. Some will have no car, some will have a relatively old used car, etc. Ideally we'd like to get person-specific characteristics for the default and use these to describe u_{i0}, then deviate all other choices from this baseline.

As in Berry (1994), define

$$\delta_j = x_j\overline{\beta} + \xi_j.$$

The probability that agent i chooses product j is

$$p_{ij}(z_i, x_j, \delta_j) = \int_\nu \frac{\exp(\delta_j + x_j(\beta^0 z_i + \beta^\nu \nu))}{\sum_k \exp(\delta_k + x_k(\beta^0 z_i + \beta^\nu \nu))} f(\nu) d\nu$$

i.e., a mixed logit. Note that the model includes a complete set of product fixed effects δ_j, which absorb the "main effects" of the product characteristics, $x_j\overline{\beta}$, as well as the common unobserved "product effects", ξ_j. Note too that if we knew $(\overline{\beta}, \beta^0, \beta^\nu)$ and had a random sample of potential buyers then we could use the estimates to predict responses to new product introductions, and to characterize the responsiveness of demand to variation in the price of product j.

A simple and appealing approach to this type of model is a two-step estimation strategy. In the first step, estimate the mixed logit model by simulated ML, including unrestricted product effects. This "first step" yields estimates of (β^0, β^ν), as well as the J product fixed effects $\{\widehat{\delta}_1, \widehat{\delta}_2, ...\widehat{\delta}_J\}$. In the second step, fit the relation

$$\widehat{\delta}_j = x_j\overline{\beta} + \xi_j + (\widehat{\delta}_j - \delta_j)$$

by IV (or a "gls" variant of IV) using instruments for price that are orthogonal to unobserved tastes for a particular brand. This identifies $\overline{\beta}$, so all elasticities are known. Note that this setup easily extends to the situation where we know the first choice and the second best alternative (as in Hastings, Kane and Staiger), provided that "not buying" is included as an option to list as the second best. (If it's not, we have to assume that the second best is always better than not buying).

Computationally, there are some difficulties in the first stage if the number of choices is large and the number of random effects is also large. One important "trick" is to concentrate the $\delta_j's$ out of the likelihood, and maximize over (β^0, β^ν).Recall from introductory econometrics class that if you have a log likelihood $\ell(\theta_1, \theta_2)$ that depends on two subsets of coefficients, θ_1 and θ_2, then the concentrated log likelihood is

$$\ell_c(\theta_1) = \max_{\theta_2} \ell(\theta_1, \theta_2) = \ell(\theta_1, \theta_2^*(\theta_1)),$$

where $\theta_2^*(\theta_1)$ is the choice for θ_2 that is likelihood maximizing, given a particular value of θ_1. In our context $\theta_1 = (\beta^0, \beta^\nu)$ and $\theta_1 = (\delta_1, \delta_2, ...\delta_J)$. We can use BLP's "contraction mapping" to get the likelihood maximizing choices for the $\delta's$ at each (β^0, β^ν). In the actual computation it helps a lot to save the choices for the $\delta's$ at each evaluation of the likelihood, and use these as "starting values" for the contraction mapping in the next evaluation. The advantage of the concentrated likelihood is that it has relatively few parameters – getting a lower-dimensional numerical optimization problem to converge is usually much easier than working with a high-dimensional problem, even if each function evaluation is more time consuming.

What do we do if we don't observe the non-buyers (as happens in applications where the product choice information comes from a survey of buyers)? In the current version of her paper, Langer is proposing an "augmented sample" approach, which essentially involves adding "fake" observations for the non-buyers. Make the following assumptions: (1) the overall fraction of non-buyers in the population, q_0 is known; (2) the observed demographic variables have a discrete distribution – i.e., $z_i \in \{z_1, ...z_P\}$ (for example, age/race/income cells); (3) the population distribution of z_i is known (i.e., the numbers $\pi_p = p(z_i = z_p)$ are known); (4) non-buyers all receive utility $u_{i0} = \epsilon_{ij}$. Under these assumptions, it is possible to augment the sample with data for the missing non-buyers. These assumptions ensure that there is no "information loss" in the augmented observations.

The method is as follows. For each "cell" of the observed $z's$ (z_p) begin by calculating the fraction of the cell who are non-buyers. (This is a straighforward calculation that depends on the relative number of observations in the sample in cell p, the relative number of observations in the population in cell p, and the overall fraction of the population who are non-buyers). Based on this fraction, calculate the number of "missing" non-buyers in the cell, and augment the sample with an appropriate number of observations, all assumed to have actually chosen not to buy. Then fit the choice model to the augmented sample by simulated ML.

In BLP (2004), a full ML approach to the "first stage" is too computationally burdensome (they have around 200 choices, allow random coefficients on a relatively large number of attributes, and relax the assumption that $u_{i0} = \epsilon_{ij}$). They instead use a method of moments approach, matching 3 sets of moments: (i) the product shares (ii) the covariances between consumer characteristics z_i and the characteristics x_j of their first choice (iii) the covariances between the characteristics x_j of each consumers first and second choices.

2. Aside: Sorting Effects in Hedonic Models

In a well-known paper, Sandra Black (QJE, 1999) compared house prices on opposite sides of school attendance zone boundaries and argued that the difference in prices revealed the willingness to pay for school quality. Assuming that houses on opposite sides of the boundary share the same neighborhood characteristics, this "boundary design" differences out shared neighborhood characteristics that typically confound a standard hedonic regression model. For concreteness, consider a simple reduced-form model like

$$p_h = a + f(X_h) + cQ_{s(h)} + e$$

where p_h is the price (or annual user cost) of house h, X_h is a vector of characteristics of the neighborhood (e.g., the crime rate, mean income, and racial composition of neighboring families), and $Q_{s(h)}$ is a measure of quality for the school district $s(h)$ that house h is assigned to. By assumption, houses on opposite sides of the boundary have the same X_h and differ only in whether the children can attend school system 1 with quality Q_1, or system 2 with quality Q_2. Thus, we can estimate

$$\widehat{c} = \frac{\frac{1}{N_1}\sum_{h\in s_1} p_h - \frac{1}{N_2}\sum_{h\in s_2} p_h}{Q_1 - Q_2}.$$

Note that this approach is easily extended to allow for observable differences in house characteristics like size or state-of-repair.

While this is a definite improvement over a simple regression approach, one may be concerned that the estimate of c suffers from what could be called "sorting bias": the observed gap in prices overstates the willingness to pay for some families, and overstates it for others. Consider family k with income y_k and utility function $U^k(y_k - p, X, Q)$ defined over other goods, characteristics X, and school quality Q. Assuming $Q_1 > Q_2$, define:

$$U^k(y_k - p_1^*, X, Q_1) = U^k(y_k - p_2, X, Q_2)$$

By construction, $p_1^* - p_2$ is family k's willingness to pay, given X and their income. Taking a first order expansion and letting $mrs^k = U_3^k(y_k - p_2, X, Q_2)/U_1^k(y_k - p_2, X, Q_2)$, note that

$$\frac{p_1^* - p_2}{Q_1 - Q_2} = mrs^k.$$

For a family that lives on the S_1 side:

$$U^k(y_k - p_1, X, Q_1) > U^k(y_k - p_2, X, Q_2) \Rightarrow \frac{p_1 - p_2}{Q_1 - Q_2} < mrs^k,$$

while for a family that lives on the S_2 side:

$$U^k(y_k - p_1, X, Q_1) < U^k(y_k - p_2, X, Q_2) \Rightarrow \frac{p_1 - p_2}{Q_1 - Q_2} > mrs^k.$$

In the special case where everyone has the same preferences, Black's procedure estimates the willingness to pay (of everyone). Otherwise, the observed price differential is set to equilibrate supply and demand for school quality.

Another problem for a simple hedonic approach is that estimates of the willingness to pay for a particular amenity represent "local average treatment effects." When there is heterogeneity in mrs^k, driven for example by variation in income levels across consumers, and the supply of the attribute in question in limited, one can recover estimates of willingness to pay from simple hedonic models that are very far from the mean willingness to pay in the population. For example (as pointed out by BFM), the variation in house prices in Pacific Heights with respect to view of the GG bridge does not reveal the mean (or even the 90th percentile) of the distribution of willingnesses to pay for a good view among all residents of San Francisco.

A final (and potentially very difficult) problem arises when one (or more) of the important amenities of a neighborhood depend on the characteristics of the people who live there (like race, income, or cognitive ability). In this case, the characteristics of the population who choose a neighborhood are endogenous. For example, consider a city with a range of elevations and assume that higher income people have a higher willingness to pay for elevation. In equilibrium, higher elevation neighborhoods will have richer families who supply higher-scoring children to local schools. This endogenous stratification can substantially magnify the observed sorting by income, race, etc., along relatively minor (or even arbitrary) dimensions like elevation. In the limit (described by theoretical models like the one in Epple and Romano, AER, March 1998) stratification is complete and there is no overlap in the characteristics of families who live in different school districts. Some evidence of this problem is apparent in the comparisons reported in BFM of the families that live on either side of a school boundary.

3. Neighborhood choice

Reference: Patrick Bayer, Fernando Ferreira, and Robert McMillan. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." Journal of Political Economy (2007) 115(4). (BFM)

BFM present an interesting application of logit demand models to the problem of neighborhood equilibrium. They are particularly concerned with trying to derive willingness to pay for "school quality" and "neighborhood quality" in the presence of heterogeneous preferences. They consider a MNL model in which household i gets utility u_{ih} from house h:

$$u_{ih} = \beta_i X_h - \alpha_i p_h - \gamma_i d_{ih} + \theta_{bz(h)} + \xi_h + \epsilon_{ih},$$

where X_h is a vector of characteristics of the house and the area (including things like the number of rooms as well as neighborhood characteristics like mean income), p_h is the price of house h (or log price), d_{ih} is a measure of the distance from house h to the head of household i's place of work, $\theta_{bz(h)}$ represents a fixed effect for the "school boundary zone" that house h is located in, ξ_h represents the unobserved "quality" of house h, and ϵ_{ih} is a EV-I error. BFM assume that the coefficients $(\beta_i, \alpha_i, \gamma_i)$ vary with some observed characteristics of the household (Z_i):

$$\beta_i = \beta + \pi_x Z_i \ ; \ \ \alpha_i = \alpha + \pi_p Z_i \ ; \ \ \gamma_i = \gamma + \pi_d Z_i.$$

(This is not exactly their notation - apologies for the confusion). Note that they do not have a real mixed logit – all heterogeneity is assumed to be captured by the interactions with Z_i. Collecting terms as in Berry and BLP, we have

$$\begin{aligned} u_{ih} &= \delta_h + \lambda_{ih} + \epsilon_{ih}, \qquad \text{where} \\ \delta_h &= \theta_{bz(h)} + \beta X_h - \alpha p_h + \xi_h \\ \lambda_{ih} &= \pi_x Z_i X_h + \pi_p Z_i \ p_h - \gamma d_{ih} - \pi_d Z_i d_{ih}. \end{aligned}$$

The probability that household i chooses house h is

$$p_{ih} = \frac{\exp(\delta_h + \lambda_{ih})}{\sum_k \exp(\delta_k + \lambda_{ik})},$$

where the "choice set" in the denonimator includes all the other houses in the local market (in their case, the 6 counties of the SF Bay area).

They follow the two-step estimation procedure described above. In the first step they estimate the parameters $(\pi_x, \pi_p, \gamma, \pi_d)$, and a full set of δ_h's (the fixed effects for each house). This is pretty computer intensive because their sample includes 27,500 houses. In the second step they estimate the mean parameters (β, α) by applying IV to:

$$\widehat{\delta}_h = \theta_{bz(h)} + \beta X_h - \alpha p_h + \xi_h + (\widehat{\delta}_h - \delta_h),$$

using characteristics of houses in neighborhoods some distance from the each house as instruments for price. (In a third step they refine the instruments – see their paper for details).

There are several aspects of this paper that could be expanded, and definitely warrant further thinking:

(a) it is unclear how the instruments can work if the boundary zones are defined precisely, because houses in given boundary zone but opposite sides of the border are almost equidistant to other neighborhoods.

(b) estimating a fixed effect for each house leads to a potential consistency problem for the first stage estimates. Under conventional asymptotics, the number of $\delta' s$ increases 1:1 with the sample size. Generally, in nonlinear models there is an "incidental parameters" problem in having the number of parameters increase with the sample size. Berry, Linton, and Pakes (ReStud, 2004) present the required conditions for BLP-related procedures to yield consistent second-stage estimates. In the BFM framework, these include a condition that the number of families increase at a certain rate relative to the number of houses that are assigned a separate fixed effect.

(c) BFM ignore the fact that many households have no children, and therefore receive no direct value from local schools. Note that the co-existence of families with and without kids in the same neighborhood requires heterogeneity in tastes to make sense!

(d) If neighborhood characteristics like racial composition are endogenously determined, it is possible that we need instruments for them. (In other words, if ξ_h is correlated with p_h, it is arguably correlated with the mean income of the neighborhood and the fraction of minority residents).

(e) BFM's specification, like the baseline BLP specification, ignores any heterogeneity in the valuation of unobserved neighborhood attributes.

Extensions

It would be extremely useful to endogenize neighborhood composition. Imagine there are 2 groups (g=1, 2) and the valuation that person i in group c places on neighborhood j has the form:

$$u_{icj} = \beta_{ic} X_j - \alpha_{ic} p_j + \gamma_c Y_{cj} + \xi_j + \epsilon_{cij},$$

where p_j represents a standardized price for a unit of housing in neighborhood j, and Y_{cj} represents the fraction of housing units held by people of group c in neighborhood j. It might be possible to use this as a "microfoundation" for a tipping model. Note the problem: this is very similar to a peer-effects model, so the usual "reflection problem" is present.

4. Nested logit and other extensions of MNL

The alternative to a mixed logit approach is to relax the independent EV-I specification for the error terms in the random utility model. Return to the baseline case (suppressing subscripts for individuals):

$$u_j = v_j + \epsilon_j, \quad j = 1...J.$$

Note that the 1st option is chosen if $v_1 + \epsilon_1 > v_k + \epsilon_k$, for all $k \neq 1$, or alternatively if $\epsilon_k < v_1 - v_k + \epsilon_1$ for all $k \neq 1$. For a general joint distribution function $F(\epsilon_1, \epsilon_2, ...\epsilon_J)$, we can write

$$\begin{aligned} p_1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{v_1 - v_2 + \epsilon_1} ... \int_{-\infty}^{v_1 - v_J + \epsilon_1} f(\epsilon_1, \epsilon_2, ...\epsilon_J) d\epsilon_J ... d\epsilon_2 d\epsilon_1 \\ &= \int_{-\infty}^{\infty} F_1(\epsilon_1, v_1 - v_2 + \epsilon_1, ...v_1 - v_J + \epsilon_1) d\epsilon_1. \end{aligned}$$

The basic idea is to find a functional form for F such that this integral can be solved easily. McFadden showed that a very convenient assumption is

$$F(\epsilon_1, \epsilon_2, ...\epsilon_J) = \exp[-H(e^{-\epsilon_1}, e^{-\epsilon_2}, ...e^{-\epsilon_J})],$$

where H is a member of the class of functions from $\Re_+^J \rightarrow \Re_+$ with 3 properties: (i) homogeneous of degree 1; (ii) $\lim_{r_j \rightarrow \infty} H(r_1, ..r_j, ..r_J) = \infty$, (iii) the first partial derivatives of H are positive, and all the distinct cross-partials of order k (e.g., $\partial^k H / \partial r_{i...} \partial r_l$ for $i...l$ all distinct) are non-positive if k is even and non-negative if k is odd. This is called the class of "generalized extreme value" distributions (GEV). Taking the partial derivative F_1, then using the facts that H is homogeneous of degree 1 and H_1 is homogeneous of degree 0, it is not too hard to establish that

$$p_1 = e^{v_1} H_1(e^{v_1}, e^{v_2}, ...e^{v_J}) / H(e^{v_1}, e^{v_2}, ...e^{v_J}).$$

(See the appendix to the paper by Arcidiacono and Miller (2007) on the reading list).

Note that if $H(r_1, ..r_j, ..r_J) = \sum_{j=1}^{J} r_j$ then $H_1 = 1$, and

$$p_1 = e^{v_1} / \sum_{j=1}^{J} e^{v_j}$$

which is the basic MNL.

The "nested logit" is obtained by partitioning $\epsilon_1, \epsilon_2, ...\epsilon_J$ into K clusters or "nests", $B_1, B_2, ...B_K$ and assuming

$$H(r_1, ...r_J) = \sum_{k=1}^{K} \left[\sum_{j \in B_k} r_j^{1/\lambda_k} \right]^{\lambda_k}, \quad 0 < \lambda_k < 1$$

(analogous to a sum of CES functions. Some authors make the inner superscript δ_k and the outer $1/\delta_k$). In this case, the probability of choice i in nest $k(i)$ is:

$$p_i = \frac{\exp(v_i / \lambda_{k(i)}) \left[\sum_{j \in B_{k(i)}} \exp(v_j / \lambda_{k(i)}) \right]^{\lambda_{k(i)} - 1}}{\sum_{k=1}^{K} [\sum_{j \in B_k} \exp(v_j / \lambda_k)]^{\lambda_k}}.$$

Note that the denominator is the same for all choices, and that the second term in the numerator is the same for any choices in the same nest. Thus, if choices a and b are in the same nest have:

$$p_a / p_b = \frac{\exp(v_a)}{\exp(v_b)},$$

as in the ordinary logit. This means that conditional on being in a nest, the choices are MNL (or conditional logit). It can also be shown the choice across nests has a logit form, where the value associated with a given nest – called the "inclusive value", depends on the v_j's for all the choices in the nest in a very simple way.

An alternative derivation of the nested logit is presented by Cardell (1997). Cardell shows that there is a family of distribution functions $C(\lambda)$ with support over the real line, indexed by $\lambda \epsilon [0,1]$, such that if $v \sim C(\lambda)$ and $\epsilon \sim EV1$, with v and ϵ independent, then $\omega \equiv v + \lambda\epsilon \sim EV1$. (In fact, Cardell shows that the relation is iff). Note that the coefficient λ goes in front of the EV-1 component of the composite random variable ω. Since both ϵ and ω are EV-1, $var(\omega) = var(\epsilon) = \pi^2/6$. Using the fact that v and ϵ are independent, we have

$$var(\omega) = var(v) + \lambda^2 var(\epsilon) = var(\epsilon)$$

implying that $var(v) = (1-\lambda^2)\pi^2/6$, which is decreasing in λ. A very convenient property of the C-family is that if $v_1 \sim C(\lambda_1)$ and $v_2 \sim C(\lambda_2)$, v_1 independent of v_2,

$$v_1 + \lambda_1 v_2 \sim C(\lambda_1 \cdot \lambda_2)$$

(the proof uses the only-if part of the previous result). Finally, note that if we start from J+1 independent r.v.'s: $v_1 \sim C(\lambda_1)$, $\epsilon_j \sim EV1$, j=1...J, we have that

$$\omega_j = v_1 + \lambda_1 \ \epsilon_j \sim EV1.$$

This gives us a set of J EV1's, each with variance $\pi^2/6$ and with $cov[\omega_i, \omega_j] = (1-\lambda^2)\pi^2/6$, implying a correlation between any pair of $(1-\lambda^2)$. Cardell shows the d.f. for $(\omega_1, ... \ \omega_J)$ is

$$F(\omega_1, ... \ \omega_J) = \exp[-\sum_{j=1}^{J} e^{-\omega_j/\lambda}]^{\lambda},$$

which is the d.f. of the nested logit.

N. Scott Cardell. "Variance Components Structures for the Extreme-Value and Logistic Distributions, with Application to Models of Heterogeneity." Econometric Theory, 13 (April 1997): 185-213.