Notes on Demand Estimation

Erica Moszkowski

September 2019

Contents

1 Introduction

The purpose of these notes is to compare different approaches to demand estimation. I find it helpful to break demand estimation down into 3 parts: model, estimation, and computation. First, this helps me understand how different models lend themselves to different estimation procedures, and what computational challenges can arise for each model. Second, I can easily keep track of behavioral/economic assumptions (which come from the model), statistical assumptions (which are necessary for estimation) and assumptions that are made to simplify (in many cases, make possible) computation. Finally, you can treat these notes as a sort of "cookbook": in different scenarios, one could imagine mixing and matching behavioral models, estimation techniques and computational tricks. I don't include all sections for every model; rather, I try to include the details I find most useful.

1.1 Attribution and References

This document is heavily based on excellent notes by Frank Pinter (frankpinter.com/demand). Like his, these notes were initially written as a study guide for the Harvard IO field exam. Other helpful references include:

- Train (2009b), a textbook on discrete choice methods (see especially Chapter 3)
- The IO chapter in the Handbook of Econometrics (Ackerberg et al., 2007)
- The appendix to "A Practitioner's Guide to Estimation of Random Coefficients Logit Models of Demand" (Nevo, 2000)

1.2 Roadmap

There are many, many models of consumer choice, lots of different estimators, and thus many possible ways to estimate demand systems. Here, I will focus on the models that help us build up a foundation for BLP. To keep all the different models straight, I think of all the incremental innovations that were added, one at a time, until we got to BLP:

1. Early models of choice: random utility maximization, the IIA axiom, and the closed-form logit share formula.
2. The shift from product space to characteristics space, leading to single-characteristic models of horizontal and vertical differentiation.
3. The multinomial logit model, which allows for multiple characteristics and estimation using maximum likelihood but is still plagued by the IIA property.
4. The pure logit model, which handles the endogeneity of unobserved characteristics.
5. The nested logit model, which allows for slightly more flexible substitution patterns than pure logit.
6. The random coefficients logit model (BLP), which allows for heterogeneity in preferences and thus escapes the IIA problem.

7. Models that add more moments and more structure to BLP, including:
   (a) adding supply-side moments and a pricing model to demand-side BLP, and
   (b) MicroBLP, which adds micro-level data to the BLP framework.

2 Background/Intellectual History

2.1 Random Utility Maximization

Agent i has utility u_{ij} from product j:

$$u_{ij} = V_j + \epsilon_{ij}$$

where V_j is fixed for everyone and ϵ_{ij} is chosen randomly for each person. What is the probability that the agent chooses a given alternative? Before we had good computers, people spent a lot of time finding distributions for which you could write down closed-form choice probabilities. For example, if ϵ_{ij} is iid Normal, we get a probit.

2.2 Independence from Irrelevant Alternatives (IIA) Axiom

The IIA axiom assumes that the ratio of any 2 choice probabilities doesn't depend on the rest of the choice set, which allows the analyst to infer choice probabilities using a binomial model. Letting $P_C(i)$ denote the probability that an agent chooses good i when the choice set is C, the binomial model implies:

$$\frac{P_C(i)}{P_C(j)} = \frac{P_{\{i,j\}}(i)}{P_{\{i,j\}}(j)}$$

Any model with this property directly imposes severe restrictions on substitution patterns, which can lead to obviously incorrect estimates of cross-price elasticites. This problem is often called the "red bus-blue bus problem," after the following thought experiment: suppose that the only options at time 1 are a train and a red bus. At time 2 we add a blue bus, which is identical to the red bus. Intuitively, the agent should be just as likely to pick the train as he was before. However, in a model in which IIA holds and all choice probabilities are positive, adding the blue bus reduces the probability that the agent chooses the train.

Why use this axiom if it makes such ridiculous predictions? As usual in economics, restrictions buy tractability. If all choice probabilities are positive, then IIA implies that choice probabilities have the following form (Luce 1959):

$$P_c(j) = \frac{w_j}{\sum_{k \in C} w_k}$$

where w_j is a positive constant weight that doesn't depend on the choice set.

It turns out that RUM and IIA are equivalent under certain conditions. Specifically, IIA is consistent with a RUM model of the form above if and only if $\epsilon_j \overset{iid}{\sim}$ Type I extreme value ($F(\epsilon) = \exp(-\exp(-\epsilon))$).

2.3 Product Space vs. Characteristics Space

So far, all the models I've written down are in "product space," that is, I have assumed individuals get utility directly from the goods they purchase. This is easy enough to understand, but it is restrictive in a number of ways. First, working in product space limits the number of products that a researcher can study at any given time. To see this, suppose I wanted to compute a matrix of cross-price elasticities for J goods. That means I'd need to estimate $\frac{J\times(J-1)}{2}$ parameters, which is on the order of J^2. To have any hope of identifying these parameters, I would need to have the same number of moments, which grows pretty quickly as J grows. Second, working in product space makes it impossible to predict how consumers will respond to any new products that might be introduced to the market.

Instead, we can imagine that consumers derive utility from characteristics of the goods that they purchase, rather than from the products themselves. From this perspective, products are nothing more than bundles of characteristics. Working in "characteristics space" means that our models can predict how demand will change when a new product enters the market, as long as we can describe that product using the characteristics we've been studying. Also, the number of parameters does not scale with the number of products in the market.

3 Pure horizontal product differentiation (Hotelling)

In the Hotelling model, products and individuals are characterized only by their locations in 1-dimensional space. Individuals prefer products that are closer to them. Utility is assumed to be quasilinear in money and subject to quadratic transportation costs:

$$u_{ij} = \bar{u} + \underbrace{(y_i - p_j)}_{\text{util from \$}} - \underbrace{\theta(\delta_j - \nu_i)^2}_{\text{transport cost}}$$

where δ_j is the location of product j, ν_i is the location of person i, and $\bar{u}$ normalizes the utility level.

Since consumers disagree on the relative values of a characteristics, we refer to preferences as "horizontal."

4 Pure vertical product differentiation

4.1 Model

4.1.1 Utility specification

In this model, everyone agrees on the relative quality of different goods, and market shares of lower-quality goods comes only from the fact that they have lower prices. Utility is:

$$u_{ij} = \bar{u} - \nu_i p_j + \delta_j$$

with $\nu_i > 0$ (ν_i is i's price sensitivity).

References: Mussa-Rosen, Gabsewicz-Thisse, Shaked-Sutton, Bresnahan (1987 JIE)

4.1.2 Required assumptions to get market shares

Order the goods by price so that good 1 is the lowest-priced good. For every good to have positive demand, ordering the goods by quality and price must result in the same ordering. In other words: higher priced goods must be of higher quality: $p_j > p_{j'} \iff \delta_j > \delta_{j'}$.

- For all agents i, $u_{i,j} < u_{i,j+1} \implies u_{i,j} < u_{i,k}$ for all $k > j+1$
- δ_j and $\frac{\delta_{j+1}-\delta_j}{p_{j+1}-p_j}$ are both increasing in j. If this were not satisfied for some good j, that good would never be purchased.

4.1.3 Market shares and price elasticities

An agent chooses good 0 if $0 > \max_{j\geq 1} -\nu_i p_j + \delta_j$. But we can use the order structure of the goods to be more specific than this. Agent i prefers item j to item k (where $p_j > p_k$) if and only if:

$$\begin{aligned} -\nu_i p_j + \delta_j &> -\nu_i p_k + \delta_k \\ \delta_j - \delta_k &> \nu_i (p_j - p_k) \\ \nu_{jk} \equiv \frac{\delta_j - \delta_k}{p_j - p_k} &> \nu_i \end{aligned}$$

Any agent with price sensitivity $\nu_i < \nu_{jk}$ prefers object j, and an agent with $\nu_i > \nu_{jk}$ prefers object k. Assume not buying any good costs 0 and has a quality of 0. This implies that an agent chooses not to buy any good at all if and only if $\nu_i > \nu_{10} = \frac{\delta_1 - 0}{p_1 - 0} = \frac{\delta_1}{p_1}$.

If $\nu_i \overset{iid}{\sim} LogNormal(\mu, \sigma)$, then $\mathbb{E}[\nu_i] = \exp[\mu + \sigma v]$ where v is standard normal. This means that agents choose good 0 if and only if

$$\exp[\sigma v + \mu] > \frac{\delta_1}{p_1}$$

Letting $\theta = (\mu, \sigma, \delta_1, \ldots, \delta_J)$ and $\psi_0(\theta) = \frac{1}{\sigma}\left[\ln(\frac{\delta_1}{p_1}) - \mu\right]$, then an agent chooses not to buy a good if and only if $v \geq \psi_0(\theta)$, which leads to a market share of $1 - F(\psi_0(\theta))$ (F is the cdf of a standard normal).

We can do the same derivation for the market shares of each good:

$$s_j(\theta) = F(\psi_{j-1}(\theta)) - F(\psi_j(\theta))$$

From here, we can compute cross-price elasticities for each pair of goods. This allows us to see that the model severely limits substitution patterns between goods:

- The cross-price elasticities for good j are only positive for goods $j+1$ and $j-1$. A pure vertical model will be rejected whenever we actually observe people switching between non-adjacent goods in response to price changes.
- Since a Normal distribution is symmetric, own-price elasticities are going to look similar for high-priced goods and for low-priced goods, even though (based on our real-world experience) we expect that own-price elasticities should be smaller for high-price/high-quality goods.

4.2 Estimation

Since we've fully specified the distribution of price sensitivity, we can easily estimate by maximum likelihood. The likelihood function is the model's predicted share raised to the power of the observed market share (or, if we have it, counts of number of purchases):

$$\mathcal{L}(\theta) = \prod_j s_j(\theta)^{\hat{s}_j}$$
$$\log \mathcal{L}(\theta) = \sum_j \hat{s}_j \ln(s_j(\theta))$$

Why does the likelihood take this form? Think back to basic probability: the probability that a ball of color j is chosen k times out of an urn (with replacement) is $\left(\frac{n_j}{n}\right)^k$, where n_j is the number of balls of color j and n is the total number of balls. Here, the model gives us a prediction for each individual's choice probability: $s_j(\theta)$. The data gives us the number of times each good was chosen: this is precisely its market share $\hat{s}_j$. When estimating using aggregate data, the likelihood will often take this form.

The limit distribution for shares is:

$$\sqrt{n}\left(\hat{s} - s(\theta)\right) \to_D \mathcal{N}\left(0, \frac{1}{n}\left[diag(s) - ss'\right]\right)$$

and you can use this to derive standard errors as usual.

4.3 Computation

We usually maximize the log likelihood function for 2 reasons:

1. Exponentiation is pretty slow relative to multiplication.
2. If some products have low observed shares, $s_j(\theta)^{\hat{s}_j}$ will be a really small number, which is hard for computers to represent with floating point numbers. Taking the log allows us to avoid dealing with tiny numbers that computers don't like.

5 Basic multinomial logit model

5.1 Model

5.1.1 RUM foundation

Let i index consumers and j index options in the choice set. Suppose each option is described by a vector of characteristics $\mathbf{x}_j$. Individual utility is given by a fixed part (which linear in characteristics) and a "random" or unobserved part:

$$u_{ij} = \underbrace{\mathbf{x}_j'\beta - \alpha p_j}_{\text{fixed/mean utility}} + \underbrace{\epsilon_{ij}}_{\text{unobserved}} \tag{1}$$

Implicit assumptions:

- Linearity is assumed for convenience, and it is restrictive in the sense that it limits possible substitution patterns we could predict. Some papers handle price differently than other characteristics, so that price sensitivity can vary by income (BLP does this).
- Characteristics do not vary from person to person. Usually this assumption is a result of data constraints. For example, with the commuting example of , we might not have data on how far individuals live from the bus stop.
- $\epsilon_{ij} \overset{iid}{\sim}$ Type 1 extreme value. Note that iid sampling is across individuals and products.

5.1.2 Choice probabilities and cross-price elasticities

Under the distributional assumption on ϵ_{ij}, choice probabilities take the multinomial logit form (see Train 2009a for the derivation):

$$s_{ij} = P(j \in \arg\max_{k \in C} u_{ik}) = \frac{\exp(\mathbf{x}_j'\beta - \alpha p_j)}{\sum_{k \in C} \exp(\mathbf{x}_k'\beta - \alpha p_k)}$$

Note that for our model of rational consumers to be internally consistent (that is, for agents to truly be maximizing utility), we have to assume that the agent knows ϵ_{ij} when making her decision. This means that the econometrician must treat ϵ_{ij} as unobservable.

Now we can compute cross-price and own-price elasticities. If α is the coefficient on price:

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

You should try to derive this.

You can see from these formulas that we have an extreme version of the IIA problem: the distribution of a consumer's preferences over products they didn't buy does not depend on the product they actually bought. More specifically:

1. Two agents who buy different products are equally likely to switch to a particular third product should the price of their product rise. As a result, *two goods j and k with the same shares have the same cross price elasticities* with any other good: cross-price elasticities are a multiple of $s_j s_k$. What you switch to doesn't depend on product characteristics at all.
    - Ariel's favorite example of why this is absurd comes from the auto setting of BLP. Both Yugos (very low quality) and Ferraris (very high quality) have low market shares. The multinomial logit model would imply that a Yugo buyer and a Ferrari buyer would have similar probabilities of switching to a Lambourghini if the price of Ferraris went up.
2. Since there is no systematic difference in the price sensitivities of consumers attracted to the different goods, own price derivatives only depends on shares $(\partial s/\partial p) = -s(1-s)$, especially for goods with small market shares. This implies that *two goods with same share must have the same markup* in a single product firm "Nash in prices" equilibrium, and once again luxury and low quality goods can easily have the same shares.

No data you have will ever fix these problems, because they are an implication of the *model, not the data.*

5.1.3 Identification

There are a couple of points I want to make here about identification in the model:

1. One of the main reasons we've assumed that utility is linear in characteristics is that it helps us gain identification. A more general way to write down multinomial logit utility would be:

$$u_{ij} = \underbrace{\delta_j - p_j}_{\text{mean utility}} + \mu + \sigma\epsilon_{ij}$$

   but in equation 1, I've replaced δ_j (a product-space representation) with $\mathbf{x}_j'\beta$. Without this restriction, the model is under-identified. With $\theta = (\delta_1, \ldots, \delta_j, \mu, \sigma)$, I have J independent product observations (since $s_0 = 1 - \sum_{j=1}^{J} s_j$) but $J+2$ parameters. The linearity assumption allows me to have $\theta = (\beta_1, \ldots, \beta_K, \mu, \sigma)$, where K is the number of characteristics (and is less than J).

2. The standard Type 1 extreme value distribution has a mean of $\gamma \approx 0.577$ and a standard deviation of $\frac{\pi^2}{6}$. That seems weird, so it seems like it would be a good idea to replace ϵ_{ij} in the utility function 1 with $\mu + \sigma\tilde{\epsilon}_{ij}$, where $\tilde{\epsilon}$ is distributed according to a standard T1EV. Then utility would be:

$$u_{ij} = \mathbf{x}_j'\beta + \mu + \sigma\tilde{\epsilon}_{ij} \tag{2}$$

Unfortunately, it turns out we can't do that. Even if we had all the data in the world, these parameters wouldn't be identified. Here's why:

- Shifting utility up by a constant μ for all options doesn't change the choice probabilities. That means we can normalize the level of utility however we want to. Normally we do this by designating an outside option (labeled $j = 0$) and normalize $u_{i0} = 0$. The outside option is usually one whose characteristics are policy-invariant: in product markets, it's usually not buying the product. This is important because in counterfactuals we normally assume that the utility of the outside option doesn't change.

- We could just divide each term in equation 2 by σ and have a utility function of the exact same form as equation 1. This means that we can only ever identify or estimate β/σ, which doesn't have an independent interpretation. However, we can use them to compute marginal rates of substitution, which we can interpret in the usual way.

5.1.4 Equivalent Variation

Suppose we wanted to estimate the welfare effect of a change in prices. From the perspective of the econometrician, consumer i's expected utility gain from a change in price at time t, assuming no other characteristics change fom time $t-1$, is

$$\mathbb{E}\left[u_{ij}^t - u_{ij}^{t-1}\right]$$

We can think back to first-year micro and compute the equivalent variation (EV). This is the change in consumer wealth that would be equivalent to a change in consumer welfare due to the price change. McFadden (1981) shows that, if an agent's marginal utility of income α_i is constant over the price region covered by the change,

$$EV = \frac{1}{\alpha_i}\mathbb{E}\left[u_{ij}^t - u_{ij}^{t-1}\right] = \frac{1}{\alpha_i}\left[\mathbb{E}\max u_{ij}^t - \mathbb{E}\max u_{ij}^{t-1}\right]$$

In the logit model, we have a nice closed form expression for the mean utility after choices have been made. See the "nice features of logit" section of the appendix.

5.2 Estimation

5.2.1 Assumptions and setup for Maximum Likelihood

Data required: prices, market shares, and characteristics across a set M of markets.

Statistical Assumptions:

- In addition to all the modeling assumptions we've made so far, we assume we can observe all the relevant characteristics. This means that the only reason our observed market shares and our model choice probabilities differ is that we have a finite sample. If our model is true, observed choices are drawn from a multinomial distribution with choice probabilities equal to market shares.

With this assumption and data, we can estimate straightforwardly by maximum likelihood:

$$\log \mathcal{L} = \sum_{m \in M} \sum_{j \in m} \hat{s}_{jm} \times \ln\left(s_{jm}(\theta)\right)$$

where, as before,

- $\hat{s}_{jm}$ is observed market share of product j in market m
- $s_{jm}(\theta)$ is the choice probability implied by the model

5.2.2 Summing up: Problems with the Logit Model

1. Too many characteristics: if there are too many characteristics, we get a too-many-parameters problem again. So suppose we limit the number of characteristics we include in the model. But now we've left out information that affects demand: that is, our left-out characteristics are now unobservables that are correlated with price.

2. Price endogeneity: If some goods are better than others in ways that the econometrician can't see (but that market participants can), then our model is misspecified. Any seller is going to take all of their product's characteristics into account when choosing the price of their good. In the data, we'll see that consumers prefer high-priced goods, without observing the characteristics that justify those high prices. This will mess up our estimated coefficients: in particular, we might find that the coefficient on price is positive rather than negative (i.e., that people like goods more when their price is higher, ceteris paribus).

3. IIA problem, as discussed in section 5.1.2

4. Overfitting: remember that the variance of the shares goes down at a rate of $\frac{1}{n}$. That means that if you have a lot of observations, you'll overfit and the data will reject the model outright.

See section 6 to show how we handle these.

6 Adding unobservables [Berry (1994) pure logit model]

6.1 Model

As discussed in section 5.2.2, the multinomial logit model has some problems. In particular, we leave out a lot of characteristics in order to avoid the too-many-parameters problem. Berry handles these by adding a vector of unobservables ξ (one element for each product $j \in J \cup \{0\}$) to the multinomial logit model, and then using instrumental variables to identify the parameters. He assumes ξ_j is vertically differentiated: everyone agrees higher ξ_j is "better."

Utility is given by:

$$u_{ij} = \underbrace{\mathbf{x}_j'\beta - \alpha p_j + \xi_j}_{\delta_j} + \epsilon_{ij}$$

Since ξ_j is correlated with p_j, we need to instrument for price (that is, we want to find a variable z_j that affects price directly while being uncorrelated with ξ_j). If we can back out δ_j from the data, we have a standard linear IV problem, which is identified under the usual exclusion and relevancy conditions.

6.1.1 Additional assumptions (above multinomial logit)

1. There must exist a sensible outside good, which we label $j = 0$, with a known market share. We normalize the mean utility of the outside good, δ_0, to be 0.

2. The market size must be large, so that observed market shares are actually close to choice probabilities.

3. We need to take a stand on which characteristics are endogenous to ξ_j. Normally we do this by making a timing assumption: we assume that ξ_j is observed by firms after characteristics are chosen, but before prices are set. This means that technically we only need to instrument for price, so all but 1 element of z_j is equal to x_j. Often we'll use more than 1 instrument for price.

6.1.2 Market shares and price elasticities

Following the same derivation as above, market shares are:

$$s_j(\theta) = \frac{\exp(\delta_j)}{\sum_{k=0}^{J}\exp(\delta_k)} = \frac{\exp(\delta_j)}{1 + \sum_{k=1}^{J}\exp(\delta_k)} \tag{3}$$

This gives us a system of J share equations in J unknowns (the δ's). If we have many consumers, sampling error in market shares is small. Berry shows this system has a unique solution.

Notice that while including the unobservables in the model helps us deal with price endogeneity, it doesn't solve the red bus-blue bus problem. Elasticities are the same as they were in the pure logit model. To fix this we really need consumer heterogeneity, which is just not present in this model.

6.2 Estimation

6.2.1 Estimation challenges and strategy

Our goal is to estimate $\theta = (\beta, \alpha)$.

Challenges:

- Since *shares are a nonlinear function of* ξ_j (since probabilities are bounded between 0 and 1), we can't use instruments for price in a linear regression of shares on characteristics.
- We can't estimate $\{\xi_j\}_{j=1}^J$ directly because then we'd have a too-many-parameter problem. Instead, we partially specify the distribution of ξ_j using a conditional moment restriction. We'll use that restriction to estimate θ.

Strategy:

1. "Invert" the demand model to find ξ as a function of θ.
2. Interact these ξs with instruments and find the θ that gets the moment condition closest to 0.

6.2.2 Estimation using the Berry inversion

Now we use the assumption that there's an outside good and that that outside good has 0 mean utility. For all j, we have:

$$\begin{aligned} \ln \frac{s_j}{s_0} &= \ln \frac{\exp(\delta_j)}{\exp(0)} \\ \ln(s_j) - \ln(s_0) &= \delta_j \end{aligned} \tag{4}$$

We can use this to estimate the δ_j's, and we'll call the estimates $\hat{\delta}_j$. Given the assumptions from the model, all we need to compute the δ's is a *transformation of the observed data*! This formula also ensures that the constraint from problem 6 is satisfied. Now the GMM problem becomes:

$$\min_{\beta} g(\hat{\delta}; \beta)' W g(\hat{\delta}; \beta)$$

which is equivalent to estimating

$$\log \frac{s_j}{s_0} = \mathbf{x}_j' \beta - \alpha p_j + \xi_j$$

by linear IV. Note that the since the LHS is actually $\hat{\delta}_j$, there is some variance there that we have to handle for standard errors.

6.2.3 Moment condition

If we have we have product-level instruments z_j, we can form a moment condition from the linear IV setup above. We can rearrange to express

$$\xi_j(\theta) = \log \frac{s_j}{s_0} - \mathbf{x}_j'\beta + \alpha p_j$$

and then partially specify the distribution of ξ_j with the conditional moment restriction:

$$\mathbb{E}\left[\xi_j | z_j\right] = 0$$

To get an unconditional moment restriction, let $h(\cdot)$ be a vector-valued function and write:

$$\mathbb{E}\left[\xi_j h(z_j)\right] = 0 \tag{5}$$

6.2.4 Choices of instruments

In general, anything that moves prices but is determined after ξ_j is fixed is a legitimate instrument. Common choices are:

- **Exogenous cost shifters.** This is a valid assumption if firms can respond to cost shifts by changing prices, but not by changing products.
- **Non-price characteristics of the same good.** This comes from our timing assumption above. Firms can first set observable characterstics, then they observe ξ_j, and then they set prices.
- **Non-price characteristics of other goods ("BLP instruments")**. BLP use the sum of characteristics of other goods by the same firm and the sum of characteristics of other goods by other firms (in the same market). If good j is produced by firm f, then the BLP instruments are:

  $$x_{jk}, \sum_{r \neq j, r \in \mathcal{F}_f} x_{rk}, \sum_{r \neq j, r \notin F_f} x_{rk}$$

  where $\mathcal{F}_f$ denotes the products produced by firm f.

  - Armstrong (2016) shows that using characteristics as instruments become loses identification power as the number of products in the market grows. The reason is that, as more products become available, prices move closer to Nash Bertrand. Essentially, the instrument relevance condition fails.
- **Prices in other markets ("Hausman instruments")**. The idea is that prices elsewhere are a proxy for underlying costs, but are independent of demand shocks in the current market. This assumption fails if, for example, there's recently been a national ad campaign, since this would affect demand in all markets.

7 Adding heterogeneity [BLP]

7.1 Model

To fix the red bus-blue bus problem, we need to allow substitution patterns to depend on individuals' characteristics. The intuition is the following: when we increase the price of one good, the consumers who leave that good have very particular preferences. In particular, they were consumers who preferred the characteristics of that good. Consequently they will tend to switch to another good with similar characteristics. If we allow utility to depend on interactions between consumer characteristics and product characteristics, we'll be able to generate exactly the kind of substitution patterns that we expect to see.

7.1.1 Utility specification

BLP add individual-specific coefficients to the Berry model of section 6. This turns the model from a "pure logit" into a "random coefficients logit."

$$u_{ijm} = \mathbf{x}'_{jm}\beta_i + \xi_{jm} + \epsilon_{ijm}$$

where β_i is a $K \times 1$ vector (β_{ik} is the kth element):

$$\beta_{ik} = \bar{\beta}_k + \mathbf{d_i}'\beta^o_k + \nu_{ik}\beta^u_k$$

Note that I have subsumed $\ln p_j$ into x_j and α into β.

Notation:

- $\bar{\beta} \in \mathbb{R}^K$ is the mean coefficient on characteristic j.
- $\mathbf{d_i} \in \mathbb{R}^R$ is a vector of observable demographics of length R.
- β^o_k is an $R \times 1$ vector of coefficients on individual demographics.
- $\nu_i \in \mathbb{R}^K$ is a vector of unobservables that perturbs each individual's coefficient away from $\bar{\beta}$. We normally assume that $\nu_{ik}\beta^u_k$ belongs to a parametric family of distributions (usually multivariate normal with a diagonal covariance matrix); then $\nu_i \sim N(0, I)$ (so that β^u_k subsumes the standard deviation).

The full utility specification is:

$$u_{ij} = \underbrace{\sum_k x_{jk}\overline{\beta}_k + \xi_j}_{=\delta_j} + \sum_k \sum_r x_{jk} d_{ir} \beta^o_{rk} + \sum_k x_{jk}\nu_{ik}\beta^u_k + \varepsilon_{ij}$$

- Notice that β^o and β^u are essentially variance terms. They scale deviations from the "mean agent".

- The second term captures the interaction between *observable product characteristics* and *observable agent characteristics.*
- The third term captures the interaction between *observable product characteristics* and *unobservable agent characteristics.*
- We are ruling out any interactions between unobservable product characteristics and any type of agent characteristics. It should be easy to see that we need this assumption for identification.

These interactions are essential for killing the IIA problem, as we'll see when we derive price elasticities and discuss identification.

7.1.2 Market shares and price elasticities

For a consumer with unobservable type vector ν_i, the logit choice probability is:

$$s_j\left(d_i, \nu_i; \beta, \delta_j\right) = \frac{\exp\left(\delta_j + \sum_k \sum_r x_{jk} d_{ir} \beta^o_{rk} + \sum_k x_{jk} \nu_{ik} \beta^u_k\right)}{1 + \sum_{q \in M(j)} \exp\left(\delta_q + \sum_k \sum_r x_{qk} d_{ir} \beta^o_{rk} + \sum_k x_{qk} \nu_{ik} \beta^u_k\right)} \quad \text{for all } j$$

where the summation in the denominator is over all goods q in the same market as j (including j itself). If we don't have any demographic information (no d_is, is the case in problem set 1) but want consumer heterogeneity, then we can integrate out over the distribution of ν_i in the population to get market shares:

$$s_j(\beta^u, \delta_j) = \int \frac{\exp\left(\delta_j + \sum_k x_{jk} \nu_{ik} \beta^u_k\right)}{1 + \sum_q \exp\left(\delta_q + \sum_k x_{qk} \nu_{ik} \beta^u_k\right)} dF_\nu\left(\nu_i\right) \quad \text{for all } j$$

However, we can't take this version of s_j directly to the data, since we can't solve the integral exactly on a computer. We use simulation-based methods to compute it.

7.1.3 Identification

We want to estimate $\theta = (\bar{\beta}, \beta^o, \beta^u)$.

Identification of β^o, β^u:

- Comes from variation in choice sets and prices across markets. If product j exists in 1 market and not another, we can observe the choices of agents similar to those who purchased j. What those agents picked in the absence of good j tells us something about substitution patterns across markets.
- From Nevo: If you observe the same markets over time, that is especially helpful. What makes the random coefficient logit respond differently to product characteristics? In other words, what pins down the substitution patterns? The answer goes back to the difference in the predictions of the two models and can be best explained with an example. Suppose we observe three products: A, B, and C. Products A and B are

very similar in their characteristics, while prod- ucts B and C have the same market shares. Suppose we observe mar- ket shares and prices in two periods, and suppose the only change is that the price of product A increases. The logit model predicts that the market shares of both products B and C should increase by the same amount. On the other hand, the random-coeféicients logit allows for the possibility that the market share of product B, the one more similar to product A, will increase by more. By observing the actual relative change in the market shares of products B and C we can dis- tinguish between the two models. Furthermore, the degree of change will allow us to identify the parameters that govern the distribution of the random coefficients.

7.2 Estimation

High-level estimation strategy:

- $\bar{\beta}$ is a "linear" parameter, in the sense that (given β^o and β^u) we can compute $(\delta_j)_{j=1}^J$ and estimate $\bar{\beta}$ by linear GMM (which has a closed form) as we did in section 6.
  - Note: being able to concentrate out linear parameters means that we can include all types of fixed effects without affecting computation time (just de-mean the data prior to estimating the linear parameters).
- (β^o, β^u) are variance terms, which means that they are nonlinear. This means that we're going to have to use nonlinear GMM and numerically optimize the GMM objective function.

7.2.1 Nested fixed point estimation algorithm

1. Simulate S values of ν_i from a standard multivariate normal distribution. You need each ν_i to be a vector of length equal to the number of product characteristics that you believe have heterogeneous coefficients. (In the problem set, we assume there are only random coefficients on 3 of the 5 characteristics, including the constant).
2. Do step 1 of GMM with $W = I$ or $W = (Z'Z)^{-1}$
   (a) Guess a value of β^u.
   (b) Given β^u, invert/solve the nonlinear system of equations to get $\hat{\delta}(\beta^u)$. BLP does this inversion by developing a contraction mapping with a unique solution. We iterate the following until it converges to some value $\hat{\delta}_j$:

$$\delta_j^{(t+1)} = \delta_j^{(t)} + \log\left(\hat{s}_j\right) - \log\left(s_j(\beta^u, \delta^{(t)})\right)$$

   Note that the Berry inversion from the pure logit model doesn't work here, since shares depend nonlinearly on β^u. However, this should look a lot like equation 4. Note that instead of integrating analytically, we simulate the integral. There are

a lot of ways to do that (importance sampling, etc), but the easiest is to just take the average of each of our simulated agents' choice probabilities:

$$s_j(\beta^u, \delta) = \frac{1}{S}\sum_{i=1}^{S} \frac{\exp\left(\delta_j + \sum_k x_{jk}\nu_{ik}\beta_k^u\right)}{1 + \sum_q \exp\left(\delta_q + \sum_k x_{qk}\nu_{ik}\beta_k^u\right)}$$

Note that this introduces simulation error, which we should have to account for in our standard errors. In practice, people don't really account for it. If you're using enough draws and good integration methods, it shouldn't be too much of an issue.
There are other ways to do this inversion (see 6).

(c) Solve the linear IV problem for $\hat{\bar{\beta}}$, using instruments Z.

i. This ends up being standard GMM with left hand side variable equal to $\hat{\delta}$. We construct Z using chosen instruments from section 6.2.4. Remember we need instruments for both the linear and nonlinear parameters.

(d) Back out $\hat{\xi}_j = \hat{\delta}_j - \mathbf{x}_j'\hat{\bar{\beta}}$

(e) Construct GMM objective function, with moment condition $g(\hat{\xi}, \beta) = \frac{1}{J}Z'\hat{\xi}$: $g(\hat{\xi}, \beta)Wg(\hat{\xi}, \beta)'$

(f) Allow your optimizer choose a new β^u. Iterate until the optimizer converges.

3. Do step 2 of GMM with W = inverse covariance of moments: $W = \left[\frac{1}{J}Z'\hat{\xi}\hat{\xi}'Z\right]^{-1}$

7.2.2 Standard errors

Since we're doing GMM, we have the usual asymptotic normality:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$$

where

$$V = \left(\Gamma' W \Gamma\right)^{-1}\Gamma' W\left(\sum_{i=1}^{3} V_i\right) W\Gamma\left(\Gamma' W\Gamma\right)^{-1}$$

Γ is the derivative of the moment condition evaluated at the true parameters (where s_{pop} are true population shares and P_{pop} is the population):

$$\Gamma = \lim_{J\to\infty} \left.\frac{\partial \mathbb{E}\left[g\left(\theta, s_{pop}, P_{pop}\right)\right]}{\partial \theta'}\right|_{\theta=\theta_0}$$

The 3 V_is reflect the 3 (independent!) sources of variation in our estimates:

1. V_1 reflects variance that arises from heterogeneity in unobserved product characteristics (the ξ_j's). This is the type of error we usually get in GMM asymptotics, so V_1 is the

standard IV-GMM covariance matrix of the moments that we'd get if we assumed we observed δ_j perfectly:

$$\begin{aligned} V_1^J &= \mathbb{E}_z \left[Z' \xi_j \left(\theta_0, s_{pop}, P_{pop} \right) \xi_j \left(\theta_0, s_{pop}, P_{pop} \right)' Z \right] \\ V_1 &= \lim_{J \to \infty} V_1^J \end{aligned}$$

Your estimate of V_1 is the sample analog of V_1^J, replacing θ_0 with $\hat{\theta}$, s_{pop} with the observed shares and P_{pop} with the observed population. In other words, it's the standard sum of squares of the moment condition, divided by the number of observations:

$$\hat{V}_1 = \frac{1}{J} Z' \hat{\xi} \hat{\xi}' Z$$

2. V_2 reflects sampling error (since we only observe a finite sample of purchasers from the population, market shares are not exactly the same as individual choice probabilities). In BLP, under the assumption that the data capture the purchasing decisions of a large percentage of the population, this term is ignored.

$$\begin{aligned} V_2^J =& \frac{J}{n} \mathbb{E} \left(\sqrt{n} \left[g \left(\theta_0, s_{pop}, P_{pop} \right) - g \left(\theta_0, \hat{s}_n, P_{pop} \right) \right] \right. \\ & \left. \times \sqrt{n} \left[g \left(\theta_0, s_{pop}, P_{pop} \right) - g \left(\theta_0, \hat{s}_n, P_{pop} \right) \right]' \right) \\ V_2 =& \lim_{J \to \infty} V_2^J \end{aligned}$$

where $\hat{s}_n$ are the observed shares.

3. V_3 reflects simulation error (which arises when we draw "agents" from an assumed distribution to simulate market shares). When simulation error (encoded in V_3) increases, then the variance of the parameter estimates increases as well. The simulation error decreases as S (the number of agents we simulate) increases, since $\sqrt{S}$ grows more slowly than $\frac{1}{S}$ shrinks.

$$\begin{aligned} V_3^J =& \frac{J}{S} \mathbb{E} (\sqrt{S} \left[g \left(\theta_0, \hat{s}_n, P_S \right) - g \left(\theta_0, \hat{s}_n, P_0 \right) \right] \\ & \times \sqrt{S} \left[g \left(\theta_0, \hat{s}_n, P_S \right) - g \left(\theta_0, \hat{s}_n, P_0 \right) \right]' | \hat{s}_n) \\ V_3 =& \lim_{J \to \infty} V_3^J \end{aligned}$$

V_3 is estimated using a Monte Carlo procedure, substituting $\hat{\theta}$ for θ_0. Specifically:

   (a) Draw a new simulated population P_{ns} (that is, new v_i's for your ns simulated agents) independently a bunch of times.

   (b) For each new population P_{ns}^t, calculate the vector of moment conditions, $g^t = Z' \xi^t (\hat{\theta}, P_{ns}^t)$.

      i. Note that to compute the moment conditions for each new "population," you'll need to run the contraction mapping to get the δ^ts conditional on that population. You'll keep $\hat{\theta} = (\hat{\beta}^u, \hat{\bar{\beta}})$ the same for all new populations.

ii. Back out the ξ^ts ($\xi^t = \delta^t - X\hat{\hat{\beta}}$) and form the moment condition g^t.

(c) Then your estimate of V_3 is

$$\hat{V}_3 = \frac{J}{S} cov(\sqrt{S} g^t)$$

Make sure you get a $J \times J$ matrix.

7.3 Computation

7.3.1 Optimization

Most optimizers work by picking a starting point and then exploring the function from there by stepping from one point to another. They usually take bigger steps at the beginning and smaller steps near the end. They decide they have reached an optimum once they've explored the parameter space a bunch, have narrowed the search down to a small area, and tiny steps no longer change the value of the objective function. Specifically, they return an optimum once the step size or the change in the objective function is below some "tolerance". You can pick the tolerance level when you call the optimizer (often the argument is called `xtol` for step sizes and `ftol` for changes in the value of the objective function). When you're running code to get your results, you should keep the tolerance very, very tight (think on the order of 10^{-14}). You can of course keep it looser when you're just testing that the code will run.

Many (including gradient descent and its many variations) use the gradient (and sometimes even the hessian) of the objective function to choose the next parameter draw. If you have an analytic form for the gradient of the objective function, providing it to the optimizer will speed up your code a lot (remember: Γ, which you'll need to compute for standard errors, is exactly what you need to compute the gradient of the GMM objective function). If you don't have an analytic form, you can provide the finite differences approximation.

Finally, remember that *your optimizer is really stupid*. It will explore the parameter space blindly, which can sometimes cause you headaches. For example, if you know that a parameter is restricted to the nonnegative domain, you need to constrain the optimizer – it doesn't know that . You can use a routine that accommodates parameter bounds, or you can enforce these yourself by setting the objective function to a super high value (remember, most optimizers are minimizers, not maximizers) whenever the suggested parameter value is outside your bounds.

7.3.2 Numerical instability and log-sum-exp

For reasons that are well beyond the scope of IO, computers have a hard time handling numbers that are either very large or very small in magnitude. To see this, try computing `exp(1000)` and `exp(-1000)` in your favorite programming language. Mathematically, we know that these are real, finite numbers. However, these numbers just require too many digits for your computer to handle. Most languages consider `exp(-1000)` to be so small that they round it to 0. Languages will handle `exp(1000)` differently: Julia returns `exp(1000)=Inf`;

Python throws a "numerical overflow error." Issues related to very large or small values are called "numerical stability problems," and they can really break your code.

In demand estimation, this tends to cause problems when you have goods with very small market share. Suppose you have 2 goods with market shares s_1 and s_2, with s_2 close to 0. Then

$$\log\left(\exp(s_1) + \exp(s_2)\right) \approx \log\left(\exp(s_1) + 0\right) \approx s_1$$

and in floating point representation, you'll get exactly s_1. You don't want that – it will force the predicted market share of s_2 to be 0. Instead, there's a trick for computing $\log\left(\exp(s_1) + \exp(s_2)\right)$ correctly. It turns out that for any finite collection of numbers $x_1, \ldots, x_n$:

$$\log\left(\sum_{i=1}^{n} \exp x_i\right) = x^* + \log\left(\sum_{i=1}^{n} \exp\left(x_i - x^*\right)\right)$$

where $x^* = \max(x_1, \ldots, x_n)$.

7.3.3 Other notes on computation

- Make sure you use the same simulated agents for the entire estimation process. If you don't do this, your GMM optimization routine will never converge.
- When you're writing and testing your code, set the number of simulated draws to be small so that your code runs faster. Then, when you think everything is ready to go, whack up the number of draws.
- There are lots of ways to structure your code. To give you some ideas, I'll give 2 examples of ways of handling computation of the denominator of market shares conditional on some parameter draw. Neither is necessarily better or worse from an aethetic point of view. The speed of the code largely depends on what your data looks like.
  - One approach is to loop over markets and pass data just from that market (call it M) to a function that computes $1 + \sum_{k \in M} \exp(\delta_j + \sum_k x_{jk}\nu_{ik}\beta_k^u)$. This approach is nice because its easy to parallelize this computation across markets. It's also a good example of "modular" code. Modular programming is a technique that emphasizes splitting up functionalities into independent, interchangeable modules (usually functions). Each function should only require exactly the arguments it needs to compute the result it is responsible for.
  - Another approach is to have a block-diagonal matrix that links each product to the other products in that market. For example, if products 1,2, and 3 are in market 1 and products 4 and 5 are in market 2, this matrix would look like:

$$\texttt{mkt} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

You can premultiply this matrix by any vector to sum up the values of that vector within a market. For example, you'll want to compute the denominator of the share formula, and here's a simple way to do it:

$$A_j = \delta_j + \sum_k x_{jk}\nu_{ik}\beta_k^u$$

$$1 + \sum_{q \in M} \exp(A) = \texttt{ones(J,1) + mkt*exp(A)}$$

where A is a stacked vector of A_js. This can be nice if you're working in a language that's really good at linear algebra and you have a smaller number of markets.

- You can usually speed up the contraction mapping by running it in log-scale. See the appendix of Nevo (2000) for more details.
- See Conlon and Gortmaker (2019) for lots more detail on computational issues.

7.3.4 MPEC formulation

It turns out we can write the linear IV problem as a constrained optimization problem (see Dube, Fox, and Su (2012)): we optimize the GMM objective function subject to a restriction that observed market shares equal model-predicted market shares.

As usual with GMM, we want the unconditional moment restriction to hold as closely as possible while making observed and predicted market shares match. First, define the sample analog of unconditional moment condition:

$$g(\delta;\beta) = \frac{1}{J}\sum_j \left(\delta_j - \sum_k x_{jk}\overline{\beta}_k\right) h(Z_j)$$

Then solve GMM problem on these moment conditions, adding the market-share system 3 as a constraint:

$$\begin{array}{ll} \min_{\delta,\beta} & g(\delta;\beta)' W g(\delta;\beta) \\ \text{s.t.} & s_j = \frac{\exp(\delta_j)}{1+\sum_k \exp(\delta_k)} \text{ for all } j \end{array} \tag{6}$$

This is a very slick expression of the problem, and it implies that we can use any constrained optimization routine to run BLP for us. In practice, I gather that it can be tricky to work with. In particular, MPEC does not allow for absorption of fixed effects, which can be problematic if you have a lot of markets. I encourage you to try it out or do some extra research on MPEC if you are interested.

8 Adding the supply side [more BLP]

Just as in homogeneous product markets, estimating supply and demand jointly (as a system of simultaneous equations) helps get more precise estimates than if we estimate the demand curve alone. And to do counterfactuals, we need a model of price setting anyway, so we might as well use it for estimation. But in order to do this, we need data on cost shifters in addition to the data on characteristics and shares we used to estimate the demand side.

8.1 Pricing model

BLP assume that equilibrium is Nash-in-prices (a.k.a. "differentiated products Bertrand"). Then they take a linear production of log marginal cost on a vector of observable cost shifters:

$$\log(mc_j) = w_j\gamma + \omega_j$$

The multi-product firm's problem is

$$\max_p \sum_j s(p_j)\left[p_j - mc_j\right]$$

$$FOC_{p_j} : \sum_j \sum_r \frac{\partial s_r}{\partial p_j}\left[p_j - mc_j\right] + s(p_j) = 0$$

Let $\Delta(p)$ be a $J \times J$ matrix encoding ownership and demand elasticities:

$$\Delta_{jr}(p) = -\frac{\partial s_r}{\partial p_j} \cdot \mathbf{1}[r \text{ and } j \text{ are produced by the same firm}]$$

Then we can write the FOCs in matrix form:

$$p = mc + \underbrace{\Delta(p)^{-1}s(\theta; p)}_{\text{markup}}$$

where $s(\theta; p)$ is the vector of predicted market shares from our demand system.

8.1.1 Intuition for Markups

For what types of goods will markups be higher? We can read this off the formula for markups:

1. Products in a sparsely populated part of characteristics space don't have much competition, and thus are able to charge higher markups.
   (a) The motivation behind the BLP instruments is precisely to capture this effect.
2. If the type of people who buy a particular good are price-insensitive, markups will be higher.
3. Markups should be higher for firms that own more products in a particular market. When some agents substitute to other products, they may substitute to other products produced by the same firm.

8.2 Estimation

Given demand side estimates, we can compute the markups, then we can use that to get marginal cost. Then we can use that to get ω_j.

For estimation, we can rearrange this further ($e_j'\Delta(p)^{-1}$ is the jth row of $\Delta(p)^{-1}$) :

$$\begin{aligned} p_j &= \exp(w_j\gamma + \omega_j) + e_j'\Delta(p)^{-1}s(\theta;p) \\ \implies \omega_j &= \log(p_j - e_j'\Delta(p)^{-1}s(p)) - w_j\gamma \end{aligned}$$

We have 2 options for how to estimate γ:

1. We can plug in our demand estimates $(\bar{\beta}, \beta^u)$ and estimate γ as a simple linear IV problem. This is very straightforward.
2. We can jointly estimate $(\bar{\beta}, \beta^u, \gamma)$ using conditional moment restrictions on ξ_j and ω_j.

When choosing between these 2 options, keep the following things in mind:

1. Joint estimation can improve the precision of estimates for the demand-side parameters because, if the model is correct, we are using more restrictions from the model to pin down the parameter values.
2. However, if the model is not correctly specified, then we are introducing bias into the estimates.
3. We are adding J more equations and a bunch of new parameters to estimate. If there are too many cost-side parameters and not enough instruments, then we will lose identification.

Remember that BLP's identifying assumption is a restriction on the conditional distribution of the unobservables. When jointly estimating demand and supply parameters, BLP assume we can use the same set of instruments for ξ_j and ω_j:

$$\mathbb{E}\left[\xi_j | Z_j\right] = \mathbb{E}\left[\omega_j | Z_j\right] = 0$$

Joint estimation is exactly the same as in BLP without the supply side, except:

1. We stack the expressions for ξ_j and ω_j and interact that long vector with a block-diagonal matrix with Z on the diagonals. The moment function g can be written:

$$g(\delta; \overline{\beta}, \gamma) = \frac{1}{J}\begin{pmatrix} \sum_j \left(\delta_j - \mathbf{x}_j'\overline{\beta}\right) h\left(Z_j\right) \\ \sum_j \left(\log\left(p_j - e_j'\Delta(p)^{-1}s(p)\right) - w_j\gamma\right) h\left(Z_j\right) \end{pmatrix}$$

2. The second-stage weight matrix for multiple-equation GMM is:

$$W = \begin{bmatrix} Z'\xi(\theta)\xi(\theta)'Z & Z'\omega(\theta)\xi(\theta)'Z \\ Z'\omega(\theta)\xi(\theta)'Z & Z'\omega(\theta)\omega(\theta)'Z \end{bmatrix}$$

3. α is now a nonlinear parameter because it enters into Δ^{-1} nonlinearly (it's in $\frac{\partial s_k}{\partial p_j}$, and then is inverted).

Of course, joint estimation is more computationally burdensome than separately estimating the parameters would be. Note that this discussion assumes β and γ are independent. If we want them to be correlated, then we need to adjust the GMM weight matrix.

8.3 Counterfactuals

8.3.1 Merger simulation

Recall that the pricing equation (based on the Nash-in-prices assumption) is

$$p = mc + \Delta(p)^{-1} s(\theta; p)$$

where $\Delta(p)^{-1}$ encodes ownership information. To do a merger simulation:

1. Get marginal costs from estimation procedure
2. Possibly adjust marginal cost for some efficiency: $\hat{mc}(1-e)$
3. Just change the ownership data in $\Delta(p)$
4. Simulate new prices by solving the pricing equation. This can be done with a fixed point procedure:
$$p^t = mc + \Delta(p^{t-1})^{-1} s(\theta; p^{t-1})$$
or with your favorite numerical solver.

8.3.2 New product introduction (ex ante)

Suppose we want to know what will happen if a new product were introduced. If we don't have data on what happens after the product enters the market, then we somehow need to construct the following:

1. What the characteristics of the new product would be. In particular, we need a model of its unobservable characteristic ξ_j. MicroBLP does this by predicting a new ξ_j based on the estimated ξ_js of other products from the same manufacturer.
2. What the price of the new product would be. This is hard to know; MicroBLP predict it based on a regression of price on product characteristics and firm dummies.
3. What the responses of competitors would be. For this we need a pricing assumption/equation, or we need to assume that other firms just don't react.

If we get to see both pre- and post- data, we can observe or estimate these 3 things.

Intuitively, product introduction should always raise welfare, since having more choice weakly raises everyone's utility. However, the logit error can cause some problems when predicting substitution patterns following new product entry. In particular, the logit error has unbounded support, so it implies that every individual's choice probability will be strictly

positive for every product. If in reality everyone agreed that the new product was terrible, any model with logit errors in the utility would overstate the market share it would get. On the other side of the coin, if an amazing product were introduced, we would understate its market share. Finally, if, as in the hotelling model, every consumer has their own ideal product, we should only see people substitute to a new product from very similar products. The logit error would cause us to predict that some people would substitute from farther-away products.

9 Adding micro data [MicroBLP]

Berry et al. (2004) had data linking consumer demographics to the products they purchase, and survey information about those consumers' second choice products. Data on second choices are particularly helpful because they help us learn about unobservable tastes $\beta^u \nu_i$, which are important drivers of substitution patterns.

MicroBLP estimates parameters by matching moments. The survey data enters through the addition of extra moments:

1. The covariance of consumer attributes and product characteristics
2. The covariance between first and second choice characteristics

Both of these covariances tell us a lot about what substitution patterns might look like under different counterfactuals. Assuming that people reported their second choices honestly, the covariance between first and second choice characteristics tell us what the consumer would do if their choice set changed (ie no longer included the good they actually bought). This allows us to pin down information about unobservable tastes (β^u).

A Properties of the Type 1 Extreme Value Distribution

A.1 Choice probability formula

For $u_{ij} = \mathbf{x}_j'\beta + \epsilon_{ij}$:

$$s_{ij} = P(j \in \arg\max_{k \in C} u_{ik}) = \frac{\exp(\mathbf{x}_j'\beta)}{\sum_{k \in C} \exp(\mathbf{x}_k'\beta)}$$

A.2 Expected value of the maximum

For $u_{ij} = \mathbf{x}_j'\beta + \epsilon_{ij}$:

$$\mathbb{E}\left[\max_j u_{ij}\right] = \sum_j \mathbb{E}\left[u_{ij} | u_{ij} \geq u_{ik} \text{ for all } k\right] \cdot s_{ij} = \log\left(\sum_j \exp\left(\mathbf{x}_j'\beta\right)\right) + c$$

References

Daniel Ackerberg, C. Lanier Benkard, Steven Berry, and Ariel Pakes. Econometric Tools for Analyzing Market Outcomes. In J.J. Heckman and E.E. Leamer, editors, *Handbook of Econometrics*, volume 6, chapter 63, pages 4171–4276. Elsevier, 2007. ISBN 9780444506313. doi: 10.1016/S1573-4412(07)06063-1. URL http://ideas.repec.org/h/eee/ecochp/6a-63.html.

Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market. *Journal of Political Economy*, 112(1):68–105, 2004. doi: 10.1086/379939. URL http://dx.doi.org/10.1086/379939.

Christopher Conlon and Jeff Gortmaker. Best Practices for Differentiated Products Demand Estimation with pyblp. 2019. URL https://pyblp.readthedocs.io.

Daniel McFadden. Econometric Models of Probabalistic Choice. In C Manski and Daniel McFadden, editors, *Structural Analysis of Discrete Data*, chapter 5, pages 198–272. MIT Press, Cambridge, MA, 1981.

Aviv Nevo. Appendix to "A Practitioner's Guide to the Estimation of Random Coefficients Logit Models of Demand" – Estimation: The Nitty Gritty. Technical report, 2000. URL http://emlab.berekeley.edu/{~}nevo.

Frank Pinter. Demand Estimation Notes. 2019. URL http://frankpinter.com/demand.

Kenneth Train. Logit. In *Discrete Choice Methods with Simulation*, chapter 3, pages 34–75. Cambridge University Press, 2nd edition, 2009a. URL https://eml.berkeley.edu/books/choice2.html.

Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition, 2009b. URL https://eml.berkeley.edu/books/choice2.html.