ECON8853 - Industrial Organization I

Lecture Notes from Julie H. Mortimer's lectures

Paul Anthony Sarkis
Boston College

Contents

Chapter 1

Overview of Demand Systems and Vertical Model

1.1 Introduction

Demand system estimation is at the core of the Industrial Organization field of economics. In fact, it is central to many economic applications such as the study of comparative statics, welfare impacts, advertising, etc. Since there are many types of markets where estimating demand is useful, there are also many approaches to estimate demand. The goal of this chapter is to present the standard and most common approaches, work out the intuition to apply these methods and discuss their advantages and disadvantages.

1.1.1 Intuition

Before diving in the theory, let's get some intuition. A demand system is the relationship between prices and goods purchased on the consumer side. Thus, IO researchers typically want to estimate the effect of products characteristics, most commonly price, on consumers' propensity to buy the products. To do this, the IO researcher needs to observe a market and its interactions. A unit of observation is therefore a market interaction such as the purchase of a good. In this unit of observation, we would ideally observe what good was purchased, its defining characteristics, its price, the state of available competing products, etc. This makes the data requirements pretty big, which is why acquiring a good enough dataset is complicated and often expensive.

1.1.2 An early example: Bresnahan (1987)

Demand system estimation is not only used for the sake of demand analysis, it can be a very useful tool to study broader industry topics. The use of demand models to talk about an industry as a whole is the main contribution of the NEIO movement (New Empirical IO). Bresnahan (1987) is one of the first empirical assessment of competition using demand estimation. The paper studies the presence of collusion in the automobile industry by looking at a dramatic price decrease event that happened in 1955.

The intuition of this paper is quite simple: assuming marginal costs stayed relatively constant during that period, Bresnahan estimates the variation in demand elasticities and compares it to a change in competition model. Using data on 85 models over 3 years, he finds that a change from a collusive to a more competitive equilibrium is consistent with the estimated change in elasticities.

1.1.3 Approaches to demand estimation

We consider different approaches to demand estimation based on three main divides:

1. Single product vs. multi-products: whether the system to estimate allows for differentiated products or a unique product.
2. Representative agent vs. heterogeneous agents: whether the system allows for consumers to have different tastes and behavior or not.
3. Product space vs. characteristic space: whether the system considers products as whole entities or as combinations of characteristics.

1.2 Single product demand

1.2.1 Representative agent

A demand system is generally thought to be a (more or less) general approximation of a true underlying demand curve, for example:

$$\ln(q) = \alpha p + X\beta + \varepsilon$$

where the ε captures any variation that is unobserved by the econometrician. Using that equation, the objective of the econometrician is to recover the underlying parameters (α, β) in order to understand the effects of prices and characteristics on demand. Note that this specification assumes a representative consumer choosing a (total) quantity q for a given price p. On the other side of the market, the firms that provide the good might also choose their prices p and characteristics x as a function of the expected demand, which is correlated with q. If this is the case, we say that we have endogeneity (from simultaneous equations, as we've seen in econometrics). Since the coverage of IV estimation in that case is more of an econometrics topic, we will assume knowledge of this in the rest of the chapter.

1.2.2 Heterogeneous agents

While the previous model is very simple, a straightforward extension would be to allow for heterogeneous agents on the market. To do that, we need to extend the previous model in two ways: (1) use a micro-founded model for individual demand that aggregates nicely and (2) estimate aggregated demand as:

$$\ln(q) = \int \gamma_i g(\gamma) \mathrm{d}\gamma + \int \alpha_i p f(\alpha) \mathrm{d}\alpha + \int \beta_i x h(\beta) \mathrm{d}\beta + \epsilon$$

where α, β and γ follow known distributions with unknown parameters to be estimated. This extension gives us some flexibility in terms of the potential interpretations. In a simple case where there are two types of agents: women and men for example, we will be able to estimate different tastes for different characteristics.

1.3 Multi-product demand

When we enter demand systems with multiple products, we need a way to differentiate those products. There are two approaches to do this: the product space approach, where a product is a nonseparable single entity; and the characteristic space approach, where a product is a combination of various characteristics.

1.3.1 Which approach to chose?

Disadvantages of the product space approach

In the product space approach, all competing products are estimated as fixed effects, meaning that, for a demand system with J products, each demand equation will have J parameters to estimate (at least). As the number of products increase, the number of parameters to estimate will become huge (J^2 in the simplest model). This issue is called the "too many parameters" problem. Note that some alternatives with groups of products will reduce this issue, while not solving it completely. A second issue with this approach is the introduction of a new good in the market, since it is not possible to estimate fixed effects when products are not yet available.

Disadvantages of the characteristic space approach

In the characteristic space approach, the data requirements are way higher than in the product space approach. This creates two main issues: first, obviously it will be hard to get all the data about all characteristics for all products; second, inevitably some characteristics will not be observed and might create endogeneity if they have an important role in demand. Moreover, we will get the same issue as in the product space approach when new dimensions are introduced (not new products, since we would be able to estimate consumers' preferences using other products).

1.3.2 Product Space approach

This course explores product space approaches only briefly since most of modern IO revolves around the characteristic space approach.

As we've seen, the product space approach considers each product as an entity on its own, meaning that all observed and unobserved features of the product enter the "utility" derived from the product. Because different consumers can derive different utilities from the same product, we might be interested in using heterogeneous agents models. In fact, in this approach, the basic breakdown between models is whether they include a representative agent or heterogeneous agents. The heterogeneous agents models use more of the available information (better estimates) but also provide a framework for distributional policy analysis (estimating the distribution of responses). While these models have been around for a long time, they still face big issues such as treatment of zeroes (when some consumers have no access to a particular good) or aggregation issues (computational). This is why much of the work that has been done in product space revolves around representative agent models, starting with the AIDS model by Deaton and Muellbauer (1980).

In a representative agent model, the general demand model is defined as:

$$\ln q_j = \alpha p_j + \beta p_{-j} + \gamma x_j + \epsilon_j$$

where p_j represents the own-price, p_{-j} is the vector of all competing products' prices and x_j is a vector of all other observed characteristics.

As we know, prices in this model are endogenous (simultaneity issue) but here we have more than one endogenous price (we have as many as products), thus estimating this model will require a very demanding IV strategy. Moreover, we can see that each product's demand function has at least J parameters, so that the number of parameters to be estimated will increase quadratically compared to the number of products: we will typically need to reduce the dimensionality of the problem to be able to estimate the demand model.

Reducing the dimensionality

There are three popular approaches to reduce the dimensionality of the product space problem. The first two approaches are useful only for for a particular set of questions, such as situations when variety is very important (e.g. in international trade they use them all the time), while the last one is the most often used in IO.

The first approach is to use a Constant Elasticity of Substitution model, such that utility is defined as:

$$u(q) = \left(\sum_k q_k^\rho\right)^{1-\rho} \Rightarrow q_j(p) = \frac{p_j^{-1/(1-\rho)}}{\sum_k p_k^{-\rho/(1-\rho)}} \cdot M$$

This approach is very efficient at reducing dimensionality since we now have only one parameter to estimete: ρ. However, it comes with the property that all goods have the same own-price elasticity and the same cross-price elasticities. While this definitely seems implausible, this approach is used in macro models and in trade.

The second approach is to use a logit demand model (not the same logit as in the characteristic approach). In this model, we have that:

$$u(q) = \sum_k (\delta_k - \ln(q_k))q_k$$

The problem with this approach is the IIA property, which implies that elasticities depend on market shares rather than "closeness" between products.

Finally, the third empirical approach is to use a model of multi-level budgeting, as in a "utility tree". This approach's intuition is to divide the set of products into small groups and sub-groups, while allowing flexible substitution within groups. To use this approach, we need to assume two properties:

- Separability: this property ensures that preferences for products coming from one group are independent of consumption of products from other groups. Formally, this implies that the utility function takes the following form:
$$u(q) = f(u_1(q^1), ..., u_n(q^n))$$
where there are n groups.

- Multi-stage budgeting: this property ensures that consumers can allocate expenditures in stages, considering only the current level of grouping and not what is inside the group.

It follows that this approach has three steps:

1. Grouping the products together in a (defensible) way.
2. Allocate expenditures to each group.
3. Allocate expenditures dynamics between groups.

Almost Ideal Demand System (AIDS)

One of the most popular model of the multi-level budgeting approach is the Almost Ideal Demand System (AIDS) model, designed by Deaton and Muellbauer.

Within a group g, demand for a product i is defined as:

$$w_i = \alpha_i + \beta_i \ln(y_g/P_g) + \sum_{j \in g} \gamma_{ij} \ln p_j + \epsilon_i$$

where w_i is total expenditure on product i, y_g/P_g is the real expenditure on group g and p_j is the price of other products in group g. As such, we can see that demand is not affected by products in other groups, except through the effect on the group g's expenditure. Note that y_g/P_g depends on the price index of group g, which can be computed in multiple ways. The two most important ways to compute it are the simple logarithmic index:

$$P_g = \sum_{j \in g} w_j \ln p_j$$

which is a weighted average of log prices of products inside group g (weights are given by the expenditure level of each product). The second way is the exact price index of Deaton and Muellbauer (based on Shepard's lemma):

$$P_g = \alpha_0 + \sum_{j \in g} \alpha_j p_j + (1/2) \sum_{j \in g} \sum_{k \neq j \in g} \gamma_{jk} \ln p_j \ln p_k$$

which is more complex to estimate (more parameters).

Then, if there is a middle level between groups, demand can be estimated with AIDS again (using price indices instead of product prices) or by a simpler log-log specification. At this level of decision-making, both approaches are the same (in the sense that they do not follow a theoretical model). A log-log specification would look like:

$$\ln q_g = \alpha_g + \beta_g \ln y + \sum_h \delta_{g,h} \ln P_h + \epsilon_g$$

Finally, the top-level is a single log-log equation, with the addition of a set of demand shifters:

$$\ln q = \alpha + \beta \ln y + \delta \ln P + Z\gamma + \epsilon$$

Issues with AIDS

All in all, AIDS works very well when products are actually grouped in certain categories. This is the reason why it is widely used in the trade and macro-consumption literatures. It can be used in IO settings where groups could be single products but it requires instrumenting for every good!

Think of the case where there are J goods, then you need to estimate J^2 elasticities or at least $J \times (J+1)/2$ if you are assuming that cross-price elasticities are symmetric. Usually, IO economists try to escape the problem by actually grouping the goods, however all the results will be dependent on your initial choice of grouping. Researchers have to think very deeply in the mechanism they use for grouping. Grouping subject to particular characteristics lead you closer to characteristic space models, but even then, how do you group when the characteristics are continuous etc.

Finally, let's look at one of the main challenges with the product space approach. Suppose a new good is introduced in a market, how can you use the previous demand estimated which did not account for the new product? This creates two issues:

1. Equilibrium effects: the introduction of a new good typically has significant effects on the pricing of the previously existing goods.
2. Extrapolation: using the previously estimated demand implies projecting demand where it is not defined.

Examples: Hausman, Leonard and Zona (1994)

The goal of the paper is to estimate demand for beer in the US in order to perform a merger analysis and test assumptions about the firms' conduct.

The market is divided hierarchically in three levels:

1. The top-level is beer against other goods. This is estimated using a log-log expenditure function:

$$\ln e_t = \beta_0 + \beta_1 \ln y_t + \beta_2 \ln P_{bt} + Z_t \delta + \varepsilon_3$$

   where e_t is total expenditures, y_t is income and P_{bt} is the beer price index.

2. Within the beer group, the middle-level are the segments: premium, popular and light. Again, a log-log functional form is assumed so that for a given segment m, the demand for the segment is:

$$\ln q_m = \beta_m \ln e_t + \sum_{m'} \sigma_{m'} \ln \pi_{m'} + \alpha_{m'} + \varepsilon_2$$

3. Within each segment, the bottom-level contains five brands. This level is assumed to follow an AIDS:

$$w_k = a_k + \sum_j a_{jk} \ln p_j + \beta_k \ln(x/P) + \varepsilon_1$$

In the bottom-level, you have to instrument the price since it is correlated with unobserved product quality (taste, etc.) and unobserved demand shocks (special events like the World Cup, etc.). Then you must find brand-level instruments: this is where the infamous Hausman instruments come in. Hausman instruments use prices in one city to instrument for prices in other cities. These instruments are typically strong, however, their relevance is very questionable. In particular, think about nation-wide advertising which was the main criticism of these instruments. They would clearly affect both cities' prices in the same way creating correlation. Most of the time you can observe these patterns but be wary of unobserved shocks that could affect the complete market. Optimally, the best instrument would be observed input costs for the brand, tax rates, etc.

1.3.3 Characteristic Space Approach

The characteristic space approach follows a different philosophy than the product space:

- Products are considered as a vector of characteristics
- Consumers' preferences (and thus their utility functions) are defined over these characteristics.
- Consumers are assumed to choose the characteristic vector (the product) that maximize their utility.
  - One consumer makes one choice to buy a single product. Allowing for buying more than one product is computationally costly and is an open area for research.
- Demand is aggregated by simply summing over consumers' choices.

Formal base model

Assume the following context:

- A consumer i is offered J alternatives. Consumer i is identified by his characteristics, v_i, similarly to the product being defined by its characteristics x_j and its price p_j.
- He must choose one option only $j \in J$, which he will do with probability P_{ij}
- Utility of an individual i for a good j is given by:

  $$U_{i,j} \equiv U(x_j, p_j, v_i, \theta) \text{ for all } j = 0, 1, \dots, J$$

  where good $j = 0$ is referred to as the outside good.

The variable (or most probably the vector) x_j contains non-price characteristics about good j, for example the size of the engine, the AC system, the dashboard software, etc.; while v_i is the vector of consumer characteristics such as income, age, etc. Finally, θ is the vector of parameters of the utility which is to be estimated. Note that in order to estimate the model correctly, you need to be sure that all consumers (or groups of consumers) in the sample have the same access to all J goods, or in other words, they face the same choice set.

Consumer i will choose good j if and only if $U_{i,j} > U_{i,k}$ for any other good k. This means that the probability of consumer i buying good j is given by:

$$\Pr\left[U_{ij} > U_{ik} \text{ for all } k \neq j\right]$$

Now assuming that utility is given by an observable component V_{ij} and an unobservable component ε_{ij}, we can rewrite the probability of buying the good as:

$$\begin{aligned}
P_{ij} &= \Pr\left[U_{ij} > U_{ik} \text{ for all } k \neq j\right] \\
&= \Pr\left[V_{ij} + \varepsilon_{ij} > V_{ik} + \varepsilon_{ik} \text{ for all } k \neq j\right] \\
&= \Pr\left[\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij} \text{ for all } k \neq j\right]
\end{aligned}$$

Finally, define $f(\varepsilon_i)$ as the J-dimensional probability density function for consumer i. Then the probability P_{ij} can be written as:

$$P_{ij} = \int \mathbb{I}\{\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij}\} f(\varepsilon_i) \partial \varepsilon_i$$

which is (as ε_i) a J dimensional integral over $f(\varepsilon_i)$.

We will see later on that some choices could make our life easier in estimating this integral (hard to do it analytically), mainly by parameterizing the error term:

- **Multivariate normal:** $\varepsilon_i \sim N(0, \Omega) \quad \longrightarrow$ multinomial probit model.
- **Type 1 Extreme-value:** $f(\varepsilon_i) = e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}} \quad \longrightarrow$ multinomial logit model.
- other variants...

We could also turn to thinking about absolute demand for good j instead of its probability, which is given by the size of the set $S_j(\theta)$ defined as:

$$S_j(\theta) \equiv \{i : U_{i,j} > U_{i,k} \text{ for all } k\}$$

Now suppose consumers are distributed following a pdf $f(v|\theta)$, we could recover the market share of good j as:

$$s_j(\mathbf{x}, \mathbf{p}|\theta) = \int_{i \in S_j(\theta)} f(v) \mathrm{d}v$$

Given a total market size of M, total demand for good j is: $M \cdot s_j(\mathbf{x}, \mathbf{p}|\theta)$

As we've seen the complete model of characteristic space relies on utility functions and error terms distributions. This is why, unsurprisingly, the functional forms of utility and errors are exactly what will differentiate the models within the characteristic space approach.

There are 5 main models of characteristic space.

Pure horizontal model

Analog to the Hotelling model (in its simplest form), there are n ice cream sellers along a beach where consumers are distributed. A consumer i has utility:

$$U_{ij} = \bar{u} - p_j + \theta(\delta_j - v_i)^2$$

where the term $(\delta_j - v_i)^2$ captures some kind of quadratic transportation cost in the distance from i's preferences (v_i) to the characteristics of product j (δ_j). In the simple Hotelling model, δ_j and v_i are location parameters (where ice cream seller j and consumer i are on the beach).

Pure vertical model

On the pure vertical, you don't need to model a distance from consumer preferences since everyone agrees on what product is better. In this case, utility is given by:

$$U_{ij} = \bar{u} - v_i p_j + \delta_j$$

The interaction between consumer characteristics and prices show that although every consumer has the same utility for the characteristics of good j (i.e. $\bar{u} + \delta_j$), consumers differ in their willingness to pay (v_i).

Logit model

In the logit model, consumers have the same taste for the goods' characteristics but are subject to an idiosyncratic shock depending on both the product and the consumer. Utility is given by:

$$U_{ij} = \delta_j - p_j + \varepsilon_{ij}$$

where ε_{ij} follows an iid extreme-value type I distribution. Analogous to the OLS, the error term allows for identification of the other parameters as long as it is not correlated with the δ_j and the price.

Pure characteristic model

This model nests both concepts of verticality and horizontality of differentiation.

Utility is given by:

$$U_{ij} = f(v_i, p_j) + \sum_k \sum_r g(x_{jk}, v_{ir}, \theta_{kr})$$

so that it could handle vertical model if $g(\cdot)$ is 0 for cross-products variables. It could also handle the horizontal model. It is a very general form and rarely used?

BLP model

The BLP model (for the authors' names Berry, Levinsohn and Pakes), is a parameterized version of the pure characteristics model. It is probably the most commonly used demand model in the empirical literature as of now, and is also very popular in non-research applications of IO. Utility is given by:

$$U_{ij} = f(v_i, p_j) + \sum_k \sum_r x_{jk} v_{ir} \theta_{kr} + \varepsilon_{ij}$$

where the ε_{ij} extreme-value Type I error term has been added.

1.4 Application: Vertical model

As we have mentioned in earlier, in a vertical model, all consumers agree on the relative quality of products (i.e. there is a clear ranking). However, people differ in their willingness to pay for such quality. The general form of utility is:

$$U_{ij} = \bar{u} - v_i p_j + \delta_j$$

From this equation, the objective is to recover the implied shares as a function of the parameters, and then estimate these parameters by comparing with the observed shares in the data.

1.4.1 Recovering the shares

In order to recover the shares, we first need to know the shape of the distribution of v_i. In this application, we assume it follows a lognormal distribution with implicit mean μ and variance σ^2.

Then, we order all goods in increasing order of price (or δ_j) but by assumption, it has to be the same order since people would buy a better alternative if it was also cheaper.

We can show that if $U_{ij} < U_{ij+1}$, then, $U_{ij} < U_{ik}$ for all $k \geq j+1$, assuming that $(\delta_{j+1} - \delta_j)/(p_{j+1} - p_j)$ is decreasing in j (the improvement in quality is not as fast as the increase in prices).

Normalize the outside good to 0, and you get that a consumer would choose the outside good if and only if:

$$0 \geq \max_{j \geq 1} U_{ij}$$

Using the property of the previous paragraph, we know that this is equivalent to:

$$0 \geq U_{i1}$$

Thus, we have that the probability that a consumer buys the outside good is:

$$P_{i0} = \Pr\left[0 \geq -v_i p_1 + \delta_1\right] = \Pr\left[v_i \geq \delta_1 / p_1\right]$$

and more generally, the share of the outside good is given by:

$$s_0(\theta) = 1 - F(\delta_1 / p_1)$$

where θ is the parameter vector containing all δs, μ and θ and $F(\cdot)$ is the cdf of the lognormal distribution.

For inside goods, we have that a consumer would choose good 1 if and only if $U_{i1} > 0$ and $U_{i1} > U_{i2}$. Thus, the probability is:

$$\begin{aligned} P_{i1} &= \Pr\left[0 \leq -v_i p_1 + \delta_1 \geq -v_i p_2 + \delta_2\right] \\ &= \Pr\left[v_i \leq \delta_1 / p_1 \text{ and } v_i \geq (\delta_2 - \delta_1)/(p_2 - p_1)\right] \end{aligned}$$

and more generally, the share of the first good is:

$$s_1(\theta) = F(\delta_1/p_1) - F((\delta_2 - \delta_1)/(p_2 - p_1))$$

Following the same argument for the following products, we get:

$$s_j(\theta) = F[(\delta_j - \delta_{j-1})/(p_j - p_{j-1})] - F[(\delta_{j+1} - \delta_j)/(p_{j+1} - p_j)] \text{ for all } j = 2, \ldots, J-1$$
$$\text{and } s_J(\theta) = F[(\delta_J - \delta_{J-1})/(p_J - p_{J-1})]$$

This gives us a system of J equations with $J + 2$ unknown parameters in θ. This means that we can only identify J parameters (the δs, provided we know/assume the parameters of the consumer characteristic distribution). Each δ will be identified using the associated market share and price.

1.4.2 Estimation

As we have seen in the previous section, our pure vertical model is not identified if the parameters of v_i's distribution are unknown. That's why, in this particular case, we need to reduce the dimension of the problem first, and then estimate via maximum likelihood.

In the vertical model, we need to "project" the product quality δ_j onto a few characteristics x_j (a K dimension vector, such that $K < J$). Formally, we have that:

$$\delta_j = \sum_{k=1}^{K} \beta_k x_{kj}$$

Now that all goods have the same K coefficients governing the value of δ, θ is effectively a $K + 2$ dimension vector, estimable using the J shares!

1.4.3 Issues with the vertical model

There is a number of issues that arise when using this model:

1. Cross-price elasticities are 0 for all goods that are further away than direct neighboring products. This cannot capture a lot of actual variation in these elasticities.
2. Own-price elasticities are not decreasing in j, thus leading to high-priced products having comparable elasticities to low-priced products, which is thought of as an undesirable property.
3. Estimating the model relies on multiple researcher assumptions:
   - The outside good has to be chosen to be the lowest or highest good.
   - The functional form of the distribution of v_i has to be assumed.
4. The error term is deterministic given the functional form of $f(v)$.

Chapter 2

Logit, Nested Logit and Multinomial Probit

2.1 Identification of Choice Models

In order to identify all choice models we are going to see in this chapter, we will need to make several specification assumptions. These assumptions will allow us to go around two main issues that arise because of the choice process. These issues can be summarized in two statements: "only differences in utility matter" and "the scale of utility is arbitrary".

2.1.1 Only Differences in Utility Matter

Recall the choice process between products as described in the previous chapter: a consumer i will choose a product j over its alternatives if the utility derived from j, denoted U_{ij} is greater than of all other alternatives. This means that the consumer will choose the product with the greatest perceived utility, and adding the same constant to all alternatives will not change the decision. This phenomenon has a particularly important implication: only parameters that capture differences across alternatives can be identified.

Product-specific constants

It is often reasonable to include an unobserved product-specific constant in the utility derived from a product j such that $U_{ij} = V_{ij} + \varepsilon_{ij}$, where

$$V_{ij} = X_j\beta + \xi_j$$

However, since only differences in utility matter, we will only be able to capture difference between ξs, not their absolute level. To see that consider a binary choice model where $\tilde{\xi_j} - \tilde{\xi_k} = \xi_j - \xi_k$, then, we would have

$$U_{ij} > U_{ik} \Leftrightarrow \tilde{U_{ij}} > \tilde{U_{ik}}$$

and we would not be able to identify which value of ξ or $\tilde{\xi}$ is the right one. That is why, in a model with J alternatives, at most $J - 1$ alternative specific constant can enter the model, and one product must have its ξ normalized to zero.

2.1.2 Irrelevance of utility scale

Just as adding a constant to the utility of all alternatives does not change the decision maker's choice, neither does multiplying each alternative's utility by a constant, i.e. the product with the highest utility would stay the same. The scale of utility and the variance of the error term are linked by definition, since multiplying U_{ij} by a constant λ for all j would imply multiplying the variance of ε_{ij} by λ^2 for all j. That is why the econometrician will have to normalize the model by normalizing the variance of the error term.

Normalization with iid errors

In the case of iid error terms, normalizing the variance is straightforward since normalizing the variance of one error term will normalize the variance of all error terms.

Normalization with heteroskedastic errors

In the case of heteroskedastic error terms, when at least one segment of the population/alternatives has a different variance than other segments, normalizing the model has to be done by normalizing the variance of one segment and estimating the variance for all others relative to that segment.

Normalization with correlated errors

Finally, when the error term is correlated across products, normalizing for scale is more of an issue, indeed, in that case, normalizing the variance for one product is not enough to normalize the variance of other products. The solution is to set the variance of one of the error differences to a number, then estimating all other variances and covariances relative to the constant.

2.2 "Plain-vanilla" Logit

2.2.1 Choice probabilities

To derive the logit model, recall the notation used in the previous chapter where each product has a mean quality level δ_j that incorporates the characteristics of the product as well as its price, and an idiosyncratic error term ε_{ij}. Thus, we have:

$$U_{ij} = \delta_j + \varepsilon_{ij}$$

In the logit model, we assume that the error term is drawn iid, from a Type-I Extreme Value (or Gumbel) distribution. The Gumbel distribution has the following laws of probability:

$$f(x) = e^{-x}e^{-e^{-x}} \text{ and } F(x) = e^{-e^{-x}}$$

Moreover, the variance of the distribution is assumed to be $\pi^2/6$; this assumption is crucial for identification since it implies normalization of the scale of utility (this is explained in the first section of this chapter). The mean of this distribution is not 0 (!) but since only differences between utility levels are important, this fact will not cause any issues. Finally, this distribution assumption allows us to have a closed-form solution for the distribution of the difference between two error terms: specifically, let $\tilde{\varepsilon}_{ijk} = \varepsilon ij - \varepsilon ik$, we have that:

$$F(\tilde{\varepsilon}_{ijk}) = \frac{\exp(\tilde{\varepsilon}_{ijk})}{1 + \exp(\tilde{\varepsilon}_{ijk})}$$

This distribution law is almost exactly the same as the Normal distribution, so that even if the extreme-value assumption gives fatter tails and slight asymmetry, we get back a Normal distribution in the differences. Nevertheless, note that the shape of the distribution is not that important and that the more important assumption is independence of errors, which requires a well-specified model to be justifiable (i.e. no missing variables, etc.). The independence assumption also requires that the unobserved part of utility for each product (ε_{ij}) is completely independent with respect to other products! In other words, taste for products cannot be correlated in an unobserved way. If this is the case, then we need to use other models that will be described later.

The probability that consumer i buys good j, given ε_{ij}, is the conditional probability that he chooses j over all other k:

$$\begin{aligned} P_{ij}|\varepsilon_{ij} = \Pr\left[U_{ij} > U_{ik} \text{ for all } j \neq k|\varepsilon_{ij}\right] &= \Pr\left[\delta_j + \varepsilon_{ij} > \delta_k + \varepsilon_{ik} \text{ for all } j \neq k\right] \\ &= \Pr\left[\varepsilon_{ik} < \varepsilon_{ij} + \delta_j - \delta_k \text{ for all } j \neq k\right] \\ &= \prod_{j \neq k} F(\varepsilon_{ij} + \delta_j - \delta_k) \text{ (by indep.)} \end{aligned}$$

We can then proceed to compute the unconditional probability by integrating over all possible values of ε_{ij} from the Gumbel distribution:

$$P_{ij} = \int \left(\prod_{k \neq j} F(\varepsilon_{ij} + \delta_j - \delta_k) \right) f(\varepsilon_{ij}) \mathrm{d}\varepsilon_{ij} = \frac{e^{\delta_j}}{\sum_{k=0}^{J} e^{\delta_k}}$$

which is called the logit choice probability, and also equivalent to the market share of product j, under the assumption that all consumers are identical in their distributions.

Some properties

This model is very practical in the sense it is very easy to set up and displays several desirable properties:

- $s_j = P_{ij} \in (0, 1)$ so that any market share can be rationalized by the model, except in extreme cases with only one product having all the sales (in those cases we would not want a logit model anyway), or some products having no sales (then we should take them out of the sample).
- As δ_j increases, reflecting higher attributed quality, the market share will increase to 1. If δ_j decreases, then the market share goes to 0. However, note that these results only hold at the limit, meaning a good will never have a market share equal to 0 or 1. If that is the case in the data, the researcher should take the product out of the dataset. This is why estimating a logit model must be done on a time-period long enough so that all goods considered have been sold at least once (think about vending machines).
- $\sum_j s_j = 1$, meaning that one alternative will be chosen, always.
- $s_j = P_{ij}$ is everywhere continuously differentiable in the characteristics δ_j and in price p_j.

The last point is useful to compute price derivatives (and elasticities):

$$\begin{aligned}
\frac{\partial s_j(\theta)}{\partial p_j} = \frac{-\alpha e_j^{\delta} \cdot (1 + \sum_k e_k^{\delta}) + \alpha e_j^{\delta} e_j^{\delta}}{(1 + \sum_k e_k^{\delta})^2} &= \frac{-\alpha e_j^{\delta} \cdot \left[1 + \sum_k e_j^{\delta} - e_j^{\delta}\right]}{(1 + \sum_k e_j^{\delta})^2} \\
&= -\alpha \cdot \frac{e_j^{\delta}}{1 + \sum_k e_j^{\delta}} \cdot \frac{1 + \sum_k e_k^{\delta} - e_j^{\delta}}{1 + \sum_k e_k^{\delta}} \\
&= -\alpha s_j(1 - s_j)
\end{aligned}$$

And using the same method, we find $\partial s_j(\theta)/\partial p_k = \alpha s_j s_k$

Characterizing mean utility

The estimation of the mean quality characteristic δ_j can be done over a set of non-price characteristics X_j. In particular, consider the following:

$$\delta_j = X_j \beta - \alpha p_j + \xi_j$$

where α and β are taste parameters for the given characteristics and ξ_j is an independent, unobserved mean utility shock. This shock is important in order to rationalize any pattern of market shares (suppose a product has better characteristics in every way but still has lower shares).

As introduced in the first section of this chapter, product-specific constants can only be estimated by normalizing the ξ of one product to a constant. Here, the choice is already made since the outside good has a zero utility by assumption. However, another issue arises by introducing this term. Iif consumers know ξ_j, firms must also know it and price their products accordingly. This introduces endogeneity in the error term with respect to prices. Thus, the econometrician needs instruments to correct for that endogeneity. Typically, these instruments include cost shifters that would not affect intrinsic demand for the good but would definitely have an effect on prices. Note that any non-price characteristic included in X_j that would also be correlated with the unobservable characteristics ξ_j needs to be instrumented as well.

2.2.2 Logit and taste variation

The logit model is very efficient and easy to set up if you need to measure systematic taste variation, that is, variation in tastes associated with observable characteristics X_j. Variations associated with idiosyncratic randomness, ε_{ij} are assumed away as type-I extreme value random variables.

In reality, the value that agents put on a particular attribute varies across these individuals. Sometimes these tastes vary in an identifiable way, for example, low-income agents might more concerned about the price than others; sometimes you might observe people with the exact same observable characteristics choosing different options. Logit models allow for identification of these taste variations only within limits: if this variation is linked with observable characteristics, then logit models can be applied; if this variation is purely random, then we will require other models.

As an example, consider the choice of households between cars, where two characteristics enter the decision: price p_j and shoulder room x_j. A household i

places value U_{ij} on buying good j such that:

$$U_{ij} = \beta_i x_j - \alpha_i p_j + \varepsilon_{ij}$$

where you can see that the parameters α, β vary across households i. Say the shoulder room taste β_i depends only on the number of members in the household m_i such that $\beta_i = \rho M_i$; this means that m_i is positively correlated with β_i. Similarly let's say the importance of price is negatively correlated to income I so that $\alpha_i = \theta / I_i$. Substituting back into the utility function we get:

$$U_{ij} = \rho M_i x_j - \theta p_j / I_i + \varepsilon_{ij}$$

which can be estimated with a standard logit model where variables are interaction between characteristics of the product and characteristics of the household.

In contrast, suppose taste variation was subject to an error term such that $\beta_i = \rho M_i + \mu_i$ where μ_i is non-observable (a random variable). Then the utility would be written as:

$$U_{ij} = \rho M_i x_j - \theta p_j / I_i + \varepsilon_{ij} + \mu_i x_j$$

and the logit would confound the error term with $\tilde{\varepsilon}_{ij} = \varepsilon_{ij} + \mu_i x_j$ which is clearly correlated across product specifications and households.

Bottomline is logit models can handle systematic taste variation but not random taste variation. For the latter, we will need more complex models, such as the BLP model studied in the next chapter.

2.2.3 Derivatives and Elasticities

The logit model presented above displays features that are not desired in the context of elasticities. These problems imply that one would run into issues while trying to expand logit results in terms of welfare analysis, antitrust analysis, etc.

First, recall the own- and cross-price derivatives and elasticities:

$$\frac{\partial s_j}{\partial p_j} = -\alpha s_j (1 - s_j); \quad \frac{\partial s_j}{\partial p_k} = \alpha s_j s_k; \quad \epsilon_{jj} = -\alpha p_j (1 - s_j); \quad \epsilon_{jk} = \alpha p_k s_k$$

These imply that:

- Two products with the same market shares will have the same markups. Indeed, from the first-order conditions, we have that:

$$p_j - mc_j = \frac{1}{|\epsilon_{jj}|} \Leftrightarrow \frac{p_j - mc_j}{p_j} = \frac{1}{\alpha(1 - s_j)}$$

  which would be the same for $s_j = s_k$. This is obviously not intuitive but also not observed in reality.
- Own-price elasticities are higher for higher priced goods. This fact comes directly from the formula of the own-price elasticity that shows a positive effect (in absolute value) from the prices. This is counter-intuitive since it would imply that people buying higher-priced items would be more price-sensitive than people buying lower-priced goods.
- Substitution between goods only depends on relative shares and not proximity of product characteristics. In fact, cross-price elasticities are given by $\frac{\partial s_j}{\partial p_k} \frac{p_k}{s_j} = \alpha p_k s_k$, which is not a function of characteristics of neither products.

2.2.4 Independence of Irrelevant Alternatives (IIA)

We have seen in the overview of the logit model that idiosyncratic errors are independently distributed across products, following a type-I extreme value distribution. The choice of the distribution turns out not to be very important, but the independence property has some implications that the researcher should know about. In fact, the independence assumption means that the unobserved utility derived from one good (ε_{ij}) is unrelated unconditionally to the unobserved utility from another good. This assumption seems to be rather restrictive in the context of goods but helped us a lot in coming up with the solution of the logit model.

The independence of the error terms turns out to have problematic implications in the realism of the choice mechanism. In fact, suppose a consumer has a choice of going to work by car (c) or taking a blue bus (bb). Say that utility derived from both models are the same, such that $P_c = P_{bb} = 1/2$, meaning the ratio of probabilities is one. Now suppose that a red bus (rb) is introduced in the market such that it is identical in every aspect except the color to the blue bus. The probability of taking the red bus should therefore be the same as taking the blue bus: $P_{rb}/P_{bb} = 1$. However, the ratio of probabilities between the car and the blue bus has not changed because of independence of irrelevant alternatives, meaning that $P_{bb}/P_c = 1$, thus $P_c = P_{bb} = P_{rb} = 1/3$ is the prediction of the logit model. In real life though, we would expect the probability to take the car to remain exactly the same since the actual problem is to either take the bus (regardless of the color) or the car, yielding $P_c = 1/2, P_{bb} = P_{rb} = 1/4$.

This IIA problem is nonetheless also a feature of the logit model when it corresponds to reality. In fact, this property allow the researcher to consider only subsets of the complete set of alternatives and still get consistent estimates, as long as for each observation, the actual choice is kept in the set. Another practical use of that property is that if the researcher is only interested in a few choices, then they do not need to include the other choices in the dataset, leaving the data research part out of the picture.

2.2.5 Consumer Surplus

For policy analysis, the researcher is often interested in measuring the change in consumer surplus that is associated with a particular event (introduction of a product, merger, etc.). Under the logit assumption, this value of the consumer surplus also takes a simple closed form that can be easily computed from the model. We know that each consumer will choose the product that yields the highest utility. In the aggregate case, all consumers have the same tastes so that in expectation (the econometrician does not observe ε_{ij}), consumer surplus is defined as the value of utility derived from the best good. Formally,

$$\mathrm{E}\left[\mathrm{CS}\right] = \mathrm{E}\left[\max_j\{U_{ij}\}\right] = \mathrm{E}\left[\max_j\{\delta_j + \varepsilon_{ij}\}\right]$$

which yields:

$$\mathrm{E}\left[\mathrm{CS}\right] = \ln\left(\sum_{j=1}^{J} e^{\delta_j}\right) + C$$

where C is an unknown that represents the fact that absolute utility cannot be measured. This value is called the logit inclusive value, or the log-sum term. As you can see, comparing two policies is easy in this setting, let one policy be denoted by the superscript 0 and the other by the superscript 1. Then,

$$\Delta\,\mathrm{E}\,[CS] = \ln\left(\sum_{\mathcal{J}^1} e^{\delta_j^1}\right) - \ln\left(\sum_{\mathcal{J}^0} e^{\delta_j^0}\right)$$

If we had transaction-level data with individual characteristics, we could perform this analysis of consumer surplus at each unit of observation and aggregate to find our results.

2.3 Nested Logit

A natural extension of the simple logit model, allowing for richer substitution patterns and a somewhat less restrictive IIA property is the nested logit model.

A nested logit model partitions the choice set in different subsets called nests such that the actual choice of one good follows from choices among nests. For example, in the simplest nested logit, with one nest, consumers first choose a nest (a category of products) as modelled in a simple logit, then within a nest, they choose a product inside the nest (again modelled as a simple logit). This sequence of logit models creates a different type of model called the nested logit. These nests are chosen by the researcher (which is not ideal) and they provide a strong structure to the model which could affect results significantly. Following the notation of Cardell (1991) and Berry (1994), we model utility as:

$$U_{ij} = \delta_j + \zeta_{ig} + (1 - \sigma)\varepsilon_{ij}$$

where the new terms are: ζ_{ig} an idiosyncratic "nest" taste shock that applies to all goods j in the nest g ; and σ a parameter of correlation in tastes within nests (if σ is high, tastes within group are very correlated and the nest structure matters, if σ is low, then the correlation in tastes within the nest is small and the nest structure is irrelevant). At first, it might seem that the σ term will not allow for the simple logit model, but in fact, the ζ_{ig} error term follows a unique distribution distribution function such that $\zeta_{ig} + (1 - \sigma)\varepsilon_{ij}$ follows an extreme-value distribution (not type-I however, we call it Generalized Extreme Value or GEV). This makes σ important to both terms, and in particular, when $\sigma \to 0$ we go to the simple logit, and $\sigma \to 1$ yields to more within-group correlation.

We know that within the same nest, the utility level $u_{ij} = \delta_j + \zeta_{ig}(1-\sigma) + \varepsilon_{ij}$, can be reduced down to ignore the effect of ζ, since it is the same across products. We end up in the same setting as the simple logit model. This implies that the conditional share of product j, or the share within the nest, is given by the same formula as the simple logit:

$$s_{j|g} = \frac{\exp(\delta_j/(1-\sigma))}{\sum_{k\in\mathcal{J}_g} \exp(\delta_k/(1-\sigma))}$$

where the denominator for this expression can be written as D_g, the total demand for products in the group g.

Meanwhile, across groups, we have that both error terms still give a type-I EV, so again, we can think about the demand for a group as we did in the logit case:

$$s_g = \frac{D_g^{(1-\sigma)}}{\sum_h D_h^{(1-\sigma)}}$$

Finally, the share of a product j is given by the product of both the share of the group containing j and the conditional share of j within the group:

$$s_j = s_{j|g} \cdot s_g = \frac{\exp(\delta_j/(1-\sigma))}{D_g^{\sigma} \cdot \left(\sum_h D_h^{(1-\sigma)}\right)}$$

As we can see, demand for product j depends on its own quality level relative to its group, the quality of the group relative to the other groups and σ, the correlation in tastes within nests.

The outside good in this model is considered as one of its own group, so that $s_{0|0} = 1$ and with the normalization of $\delta_0 = 0$ and $D_0 = 1$, we get:

$$s_0 = 1 \cdot s_0 = \frac{1}{\sum_h D_h^{(1-\sigma)}}$$

This analytical derivation is the same when we extend the model to have nests within nests and more.

2.3.1 IIA and substitution patterns

The nest structure of this model satisfies two properties:

1. For any two alternatives that are in the same nest, the ratio of probabilities is independent of the attributes and/or existence of other products in the nest or in other nests.

   $$s_j/s_k = s_{j|g}/s_{k|g} = \frac{\exp(\delta_j/(1-\sigma))}{\exp(\delta_k/(1-\sigma))}$$

   Equivalently, we say that IIA holds within the nest.

2. For any two alternatives that are in different nests, the ratio of probabilities depends on the attributes and/or existence of other products in the two nests.

   $$s_j/s_k = s_{j|g}/s_{k|h} \cdot s_g/s_h = \frac{\exp(\delta_j/(1-\sigma))}{\exp(\delta_k/(1-\sigma))} \cdot \frac{D_g^{1-\sigma}}{D_h^{1-\sigma}}$$

   Equivalently, we say that IIA does not hold across nests.

2.3.2 Research considerations

As we have seen, the nested logit is a fairly simple extension on the simple logit case, which relies on relaxation of the IIA property across groups (or nests). The model is often told in a narrative of sequential choice: consumers first choose a group g (s_g), then a product (or group) j within group g ($s_{j|g}$).

The results of the model depend very strongly on the ex ante nesting structure chosen by the researcher. Therefore, it is very important (and hard) to understand what the appropriate structure should be. The sequential narrative is supposed to help the researcher come up with an accurate structure but it might be unhelpful at times.

Identification comes in different flavors:

- Parameters associated with characteristics and prices are identified within group by variation in these exact characteristics.
- The correlation in tastes parameter (σ) is identified by variation in

2.4 Multinomial Probit

2.4.1 Intuition

The two previous models covered in this chapter had two main issues: they cannot handle random taste variation and they imply restrictive substitution patterns due to IIA (even though the nested logit solves this issue in some way). The multinomial probit on the other side, is not affected by these two issues at all. In essence, the MNP approach allows for correlation between choices thanks to a flexible covariance matrix of the errors.

2.4.2 Choice probabilities

As in the general case, we start with writing the individual (unconditional) probability of choosing product j:

$$\begin{aligned}
P_{ij} &= \mathrm{E}\left[\Pr\left[U_{ij} > U_{ik} \text{ for all } j \neq k | \varepsilon_{ij}\right]\right] \\
&= \mathrm{E}\left[\Pr\left[\delta_j + \varepsilon_{ij} > \delta_k + \varepsilon_{ik} \text{ for all } j \neq k | \varepsilon_{ij}\right]\right] \\
&= \mathrm{E}\left[\Pr\left[\varepsilon_{ik} < \varepsilon_{ij} + \delta_j - \delta_k \text{ for all } j \neq k | \varepsilon_{ij}\right]\right] \\
&= \int \prod_{j \neq k} F(\varepsilon_{ij} + \delta_j - \delta_k) f(\varepsilon_{ij}) \mathrm{d}\varepsilon_{ij}
\end{aligned}$$

But in the multinomial probit case, the law of probability for ε_{ij} depends on all the other error terms! Thus we have correlation between choices, that we estimate.

Formally, we have:

$$\varepsilon_i \sim N(\mu, \Omega)$$

where ε_i is the vector of all product error terms. This means that Ω is a $(J+1)^2$-elements symmetric matrix. In practice, we restrict μ to be 0 (especially if we estimate ξs), but following the first section of this chapter we will also need to restrict the variance of the error terms heavily. In fact, as mentioned in that section, in the case of iid but correlated error terms, we need to normalize one diagonal term of the error-differences distribution and estimate the other terms relative to that term.

2.4.3 Applications

The MNP approach is interesting in most cases where choice correlations are important, but is limited by the fact that data requirements get huge when we increase the number of products.

2.5 Estimation Strategy for Product-level data

The general estimation strategy that applies to logit models with product-level data is detailed in the following steps:

1. Assume that data is drawn from markets with large n.
2. Assume that observed market shares are measured without errors.
3. For each θ (vector of parameters), there exists a unique ξ such that the model shares and observed shares are equal (J equations, J unknowns).
4. We invert the model to get ξ as a function of the parameters. How this step is performed depends on the functional form of the model.
5. Using ξ, we can create the moments of the model, estimating them by GMM.

2.5.1 Inversion in the logit case

We need to express the quality unobservable term ξ_j in terms of all observable characteristics. First, recall that:

$$\delta_j = X_j\beta - \alpha p_j + \xi_j$$

but in this equation, δ_j is unknown (there is no such thing as a perceived mean quality level). Nevertheless, we can use the market shares formulas that link the δ_j to the observed market shares s_j (observed without errors). Thus,

$$s_j = \frac{\exp(\delta_j)}{1 + \sum_k \exp(\delta_k)}$$

but again, we run into a problem since it is expressed as a function of other δ_k: we could solve a system of equations, or simplify the model using the normalized good. Indeed,

$$\frac{s_j}{s_0} = \exp(\delta_j) \Leftrightarrow \ln(s_j) - \ln(s_0) = \delta_j \Leftrightarrow \ln(s_j) - \ln(s_0) = X_j\beta - \alpha p_j + \xi_j$$

which in turn yields the inversion equation:

$$\ln(s_j) - \ln(s_0) + \alpha p_j - X_j\beta = \xi_j$$

2.5.2 Inversion in the nested-logit case

In the same way as in the simple logit model, our main objective is to get the link between δ_j and observables so that we can identify ξ_j. In this case:

$$s_j = \frac{\exp(\delta_j/(1-\sigma))}{D_g^{\sigma} \cdot \left(\sum_h D_h^{(1-\sigma)}\right)} \text{ and } s_0 = \frac{1}{\sum_h D_h^{(1-\sigma)}}$$

so that

$$\ln(s_j) - \ln(s_0) = \frac{1}{1-\sigma} \cdot \delta_j - \sigma \ln(D_g)$$

We turn into a new problem caused by the two-level structure, which is D_g is not parameterized. Again, we solve this issue by using the normalized good:

$$s_g/s_0 = D_g^{(1-\sigma)} \Leftrightarrow \ln(s_g) - \ln(s_0) = (1-\sigma) \cdot \ln(D_g)$$

which we can plug back in the previous equation to get:

$$\ln(s_j) - \ln(s_0) = \frac{1}{1-\sigma} \cdot \delta_j - \frac{\sigma}{1-\sigma}[\ln(s_g) - \ln(s_0)]$$

$$\Leftrightarrow (1-\sigma) \cdot [\ln(s_j) - \ln(s_0)] = \delta_j - \sigma[\ln(s_g) - \ln(s_0)]$$

$$\Leftrightarrow (1-\sigma) \cdot \ln(s_j) + \sigma \ln(s_g) - \ln(s_0) = \delta_j$$

Finally, using the fact that $s_g = s_j/s_{j|g} \Leftrightarrow \ln(s_g) = \ln(s_j) - \ln(s_{j|g})$, we can get:

$$\Leftrightarrow \ln(s_j) - \sigma \ln(s_{j|g}) - \ln(s_0) - X_j\beta + \alpha p_j = \xi_j$$

2.5.3 Endogeneity issues

Simultaneity

Regardless of the model we use, we will have to deal with endogeneity issues regarding many dimensions, such as prices, observable characteristics or market shares.

In fact, in both models price was correlated to the unobservable term as firms make their decision based on demand which includes this term; moreover, in the nested logit case, we also have to deal with endogeneity in the within-share term. In general, anything that is correlated with the unobservable quality term ξ will have to be instrumented.

Measurement errors

Measurement errors in observed prices, characteristics or quantities may also create difficulties for the estimation procedure outlined above. Prices enter linearly in the estimation and are already endogenous so that these errors do not create too important biases. However, when these errors are present in market shares or quantities data then this issue becomes more important. Indeed, market share data is used to invert the model and get the unobservable term as a function of observed data. If shares include measurement errors, the non-linear inversion will accentuate the errors in a non-tractable way, creating a lot of issues in estimation. In practice, maximum likelihood will not have this kind of issue, but will make IV strategies harder.

Identification

For identification, we need to satisfy orthogonality conditions between ξ_j and the covariates (product characteristics x_j and price p_j). As discussed earlier, price will for sure be endogenous, so we need at least one instrument for this. Formally, we need a set of instruments w such that $\mathrm{E}\left[\xi | x, w\right] = 0$.

In practice, the instruments for price will include "cost-shifters" (i.e. variables that affect price without affecting demand).

2.5.4 Adding supply-side restrictions

It is also possible to get more orthogonality conditions by adding a supply-side to the demand system. Essentially, this means adding the firms' decisions about prices as moment conditions. To do this, we have to assume two things: (1) the firms' cost functions and (2) the competition model. Then, using the FOCs of the firms, one can recover extra moments to use in combination of the ξ moments to estimate the demand system by GMM.

Chapter 3

Mixed Logit (BLP)

For now we have seen three types of discrete-choice models and their applications to demand estimation for differentiated products: the simple logit model, the nested logit model and the multinomial probit model. The first two models, although quite useful and fairly simple to estimate were limited in three dimensions: they do not allow for random taste variation (2.2.2), they have restricted substitution patterns (2.2.4), they do not allow for correlation over time. Mixed logit models (which contains BLP) are highly flexible models that can deal with the previously mentioned issues.

Mixed logit models are a class of models that encompasses all types of models where the market shares are computed as integrals over simple logit functional form. We'll see in detail later what that means but intuitively, you should think of the model as everyone having her/his own logit model of demand, and aggregate demand would be computed by integrating over consumer attributes. In particular, IO economists are most interested in the BLP (for [**?**]) model and extensions of it like [**?**].

We'll first cover the basics of mixed logit and random coefficients, before talking in depth about estimation techniques as found in BLP and DFS.

3.1 Model

Generally, we write utility derived from consumption of a good j by consumer i as the function $u_{ij}(X_j, p_j, \xi_j, \nu_i, \theta)$, which is a function of product j observable characteristics (X_j of dimension $L-1$), price (p_j), unobservable characteristics (ξ_j) and consumer characteristics (observable z_i and unobservables ν_i), all entering the utility function through a vector of parameters θ.

As a simpler case, define utility as a linear function of those parameters such that:

$$U_{ij} = \sum_{l=1}^{L} x_{jl}\beta_{il} + \xi_j + \varepsilon_{ij}$$

$$\text{where } \beta_{il} = \lambda_l + \beta_l^o z_i + \beta_l^u \nu_i$$

Notice the main differences with logit models as we know them: first, price is not separated but included in observable characteristics of product j because of the second point, that β_i is a coefficient dependent on consumer characteristics. This means that different consumers will have different tastes in the same characteristics. For example, some person might be interested in the price of a phone while someone would disregard this and focus only on memory, another one on camera quality, etc.

We can rewrite the utility function by plugging in the definition of β:

$$\begin{aligned} U_{ij} &= \sum_{l=1}^{L} x_{jl}\left(\lambda_l + \beta_l^o z_i + \beta_l^u \nu_i\right) + \xi_j + \varepsilon_{ij} \\ &= \underbrace{\sum_{l=1}^{L} x_{jl}\lambda_l + \xi_j}_{\delta_j\text{: product mean utility}} + \sum_{l=1}^{L} x_{jl}\beta_l^o z_i + \sum_{l=1}^{L} x_{jl}\beta_l^u \nu_i + \varepsilon_{ij} \end{aligned}$$

There are five elements in this utility function:

- Observed ($x_j\lambda$) and unobserved (ξ_j) product quality.
- Observed ($x_j\beta^o z_i$) and unobserved ($x_j\beta^u\nu_i$) consumer-product interactions.
- A type-I EV iid error term (ε_{ij}).

The main features of the mixed logit model rely on the consumer-product interactions. Because they are not restricted to take the logit form, they will display more reasonable subsitution patterns when aggregated. With these interactions, products will be close in terms of the characteristics of the consumers who buy the product. For example, consider an auto market following a price increase for the BMW 7 series (very expensive and luxurious car). In the logit model, this would create substitution to all other cars based on their market shares, meaning that a small Toyota Echo would benefit from the price increase. In the nested logit, provided we defined nests correctly, only cars within the same nest would be affected (luxury cars). Already, this makes more sense, but in the mixed logit model, we use parameters estimated from price variation to figure out what happened to the people who were buying BMW 7 series.

Aggregate vs. micro data

We have seen that the general form of the mixed logit model includes both observed and unobserved consumer characteristics. In general however, market data does not come with exact description of consumer characteristics for each interaction. At best, we might have aggregate consumer data (about a geographical region, a point in time, ...) but most often we cannot observe any characteristic.

When we do observe consumer characteristics, we can use them in z_i to estimate the model. A particularly influential paper using this type of data is the MicroBLP paper. When such data is unobserved, we work with just the ν_i part of the model, as in the original BLP paper. For the rest of this section, we assume that we do not observe z_i.

3.1.1 Building the shares

Recall the utility function from the previous section (without z_i by assumption):

$$U_{ij} = \delta_j + \sum_{l=1}^{L} x_{jl} \beta_l^u \nu_i + \varepsilon_{ij}$$

Using the fact the error term is a type-I EV as in the simple logit, we can integrate it and get logit shares, but only conditional on ν_i! Formally, we get:

$$P_{ij}|\nu_i = \frac{\exp(\delta_j + x_j\beta\nu_i)}{\sum_k \exp(\delta_k + x_k\beta\nu_i)}$$

$$\Rightarrow P_{ij} = \int \frac{\exp(\delta_j + x_j\beta\nu_i)}{\sum_k \exp(\delta_k + x_k\beta\nu_i)} f(\nu_i)\mathrm{d}\nu_i$$

Latent-class models

In the simplest case, consider that there are only two types of people, such that only two β_i are possible, say β_H and β_L. There are also two probabilities associated with the types. Thus the shares become:

$$P_{ij} = \frac{\exp(\delta_j + x_j\beta_H)}{\sum_k \exp(\delta_k + x_k\beta_H)} \cdot p_H + \frac{\exp(\delta_j + x_j\beta_L)}{\sum_k \exp(\delta_k + x_k\beta_L)} \cdot (1 - p_H)$$

3.2 BLP algorithm

Consider the situation where data is available only on the aggregate product-level, and no consumer data is observable. We will work out the estimation of demand following four steps:

1. Work out the aggregate shares conditional on both δ (mean utility) and β (taste variation).
2. Invert the shares to get ξ.
3. Estimate the model using the method of moments.
4. Repeat until convergence of all parameters.

To help understanding the details of the following steps, you should keep in mind a quick summary of what it to be done. As always, we are interested in the parameters of utility (that affect demand), thus λ and β^u (that we we now call β only). These parameters are estimated using the interaction of the unobservable product characteristic ξ and adequate instruments. Until now, nothing should surprise you as we follow the same steps as described the GMM estimation strategy for logit and nested-logit models. However, this time it is different since to get ξ, you will need the parameters you are looking for (this is the main difference of BLP). That is why you use starting values for those, and iterate until you find the parameters that are stable through estimation (this is a fixed point problem).

Step 1: Conditional aggregate shares

Recall that in the simple and nested logit models we studied, the probability P_{ij} that a consumer i buys product j was equal to the market share since they did not differ from the representative consumer in any way. This time, we now that two different consumers will have different probabilities of buying the product. Consequently, the market share is going to be the integral of consumers individual probabilities over their characteristics (here ν_i since we do not observe any consumer characteristics).

Recall that we computed the market shares as a function of product characteristics:

$$s_j(\delta, \beta) = \int \frac{\exp(\delta_j + X_j \nu_i \beta^u)}{1 + \sum_k \exp(\delta_k + X_k \nu_i \beta^u)} f(\nu) \mathrm{d}\nu$$

The issue with this integral is that it cannot be solved analytically (ν_i is usually a multivariate normal distribution, which makes the market share a multidimensional integral); we can approximate it by taking the sample average over a set of ns draws in a simulation. This yields:

$$\hat{s}_j^{ns}(\delta, \beta) = \frac{1}{ns} \sum_i \frac{\exp(\delta_j + X_j \nu_i \beta^u)}{1 + \sum_k \exp(\delta_k + X_k \nu_i \beta^u)}$$

which is a function of parameters that we want to estimate (δ and β). We will see that δ will be computed, and β will be estimated by doing this step multiple times and finding the best one. This is why you should keep the simulated ν_i for the whole exercise, else it is never going to converge.

Note that integration over the distribution of ν_i is a different problem than ε_{ij} for which we can use the form of the type-I EV distribution to help us with analysis. In the case of ν_i, we assume a multivariate normal distribution (across multiple consumer unobserved characteristics). If the distribution (not observations, but at least the pdf) is observed, then we can draw from the said distribution.

Moreover, notice that the use of a finite number of draws in the simulation will create a new source of errors within our estimation routine. Enough simulation draws should help tampering this issue, although finding the good number of draws is not an exact science, but more like a tradeoff between computation speed and errors. Overall, numerical evaluation of integrals is a particular topic that is deep enough to think about it carefully.

Step 2: BLP inversion

We now need to recover the product unobservable term ξ_j. In the same way as we did in the simple logit models, we already have the link between δ and ξ, but δ is "buried" in a nonlinear fashion into the the market shares: we need to invert the market shares to get delta, thus ξ, as a function of the shares (rather than the opposite). Doing that requires a special trick that is at the very core of[**?**] contribution.

Their trick is to use the fact that the following system:

$$\delta_j^k(\beta) = \delta_j^{k-1}(\beta) + \ln(s_j) - \ln(\hat{s}_j^{ns}(\delta^{k-1}, \beta))$$

is a contraction mapping. To see it, understand that s_j is the observed market share, the exponent k represents the iteration process. In other words, by iterating over this function, the δ values will converge to the true value, $\delta^*(\beta, s, \hat{s})$, where $s, \hat{s}$ are respectively the vectors of observed and estimated shares conditional on β. Finally, we can write:

$$\xi(\beta, s, \hat{s}) = \delta^*(\beta, s, \hat{s}) - X_j\lambda$$

and use this form to construct the moments.

Step 3: Constructing the moments

Now that we have recovered ξ as a function of β (and λ through β since there is a way to write λ as a function of β), we can construct the moment conditions for demand estimation. To do this, we can go the OLS route if no component of x is endogenous but most probably we will go the IV route, using w as the instrument matrix. The moment condition would therefore be:

$$\mathrm{E}\left[\xi_j(\beta) w_j\right] = 0$$

As always in GMM, we want to select β such that the average analog to the moment equation is the closest possible to 0.

Step 4: Algorithm iteration

The first three steps were performed for a given β, thus we now need to find the best β, using a nonlinear search over β.

3.2.1 Identification

In order to fully identify the model, we need four sources of variation:

1. Choice set variation:
2. Product characteristics variation:
3. Consumer characteristics variation:
4. Functional form:

3.2.2 Helping estimation

Adding data

Adding a supply side

3.3 DFS algorithm

Chapter 4

Product Availability

Chapter 5

Entry Models

5.1 Introduction

5.2 Static and exogenous models

The main question of the book from [?] revolves around how do markets evolve to be less or more concentrated? Moreover, Sutton tries to explain why advertising is higher in concentrated industries. In order to explore the answers to these questions, he looks at different market structures and analyzes what makes these structure coherent with the data.

5.2.1 Perfect competition

Consider a market with perfect competition (price-taking assumption), exogenous sunk costs and free entry. The minimum efficient level of production (in the long run) is where $p = \min AC$.

This assumption means that, in the long run, firms will produce only the quantity satisfying this assumption, not more, not less. This implies that the number of firms in a market only depends on the size of the market, M. In particular, if you denote the quantity produced by firms at the optimum as q^*, you will have that the optimal number of firms n^* is given simply by:

$$n^* = M/q^*$$

As $M \to \infty$, we also have $n^* \to \infty$, thus the concentration ratio will tend to 0.

The issue with this simplistic model is that it might not hold in many settings where (1) competition might be imperfect or (2) sunk costs might be endogenous. As an example consider the increase of population in the United States since the creation of Pepsi and Coca-Cola. Even though the population has more than doubled, their respective market shares are still about 30% and 60%, which clearly disproves the previous model. For that reason he explores both directions.

5.2.2 Imperfect competition

Now suppose you allow for imperfect competition. For that purpose, we consider a two-stage game where in the first stage, firms choose to enter a market and incur a cost of A; then in the second stage, firms play the market game.

Moreover, assume that all M consumers have an income of 1 to spend exclusively on the good produced by the firms, such that in equilibrium:

$$p \cdot Q = M$$

Finally, suppose the marginal cost of production is $c > 0$.

Bertrand competition

Cournot competition

In the Cournot solution, assuming N firms enter, a firm i will choose its quantity q_i such that:

$$q_i \in \arg\max_{q_i} \left[M \cdot \left(\sum_{j=1}^{N} q_j \right)^{-1} - c \right] \cdot q_i$$

which yields the following FOCs:

$$\left[M \cdot \left(\sum_{j=1}^{N} q_j\right)^{-1} - c\right] - q_i \cdot M \cdot \left(\sum_{j=1}^{N} q_j\right)^{-2} = 0 \text{ for all } i = 1, 2, \ldots, N$$

$$\Leftrightarrow \left[M \cdot Q^{-1} - c\right] - q_i \cdot M \cdot Q^{-2} = 0 \text{ for all } i = 1, 2, \ldots, N$$

If we add all N conditions together, we get:

$$N \cdot \left[M \cdot Q^{-1} - c\right] - Q \cdot M \cdot Q^{-2} = 0$$

$$\Leftrightarrow N \cdot M \cdot Q^{-1} - N \cdot c - M \cdot Q^{-1} = 0$$

$$\Leftrightarrow \frac{N-1}{N} \cdot M \cdot Q^{-1} = c$$

Finally recall that $M \cdot Q^{-1}$ is actually the equilibrium price, meaning that we have: $p = \frac{N}{N-1}c = (1 + \frac{1}{N-1})c$. Consequently, we have:

$$Q = \frac{N-1}{N} \cdot \frac{M}{c}; q_i = \frac{N-1}{N^2} \cdot \frac{M}{c}; \Pi(N) = \frac{c}{N-1} \cdot \frac{N-1}{N^2} \cdot \frac{M}{c} = \frac{M}{N^2}$$

Hotelling/Salop competition

5.3 Static and endogenous models

5.3.1 Sutton's view

5.3.2 Bresnahan and Reiss (1991)

In a realistic research setting, it is hard to observe strategic variables on a market level (prices, costs, advertising expenses, for all firms). This is the reason why [?] innovated on an estimation using mainly superficial market observables: the number of firms (but not the shares) and other market characteristics. The main assumptions of this model rely on a static Sutton model with symmetric firms and free entry.

Behavioral model

As in the Sutton-type of models, we solve it backwards, assuming the fixed costs of entry are incurred as sunk in the second period, when firms are choosing production.

Demand in the market m is given by:

$$Q_m = d(Z_m, p) \cdot S(Y_m)$$

where $d(\cdot)$ represents the demand function of a representative consumer and $S(\cdot)$ is the total number of consumers that would buy the product. Note that this demand specification has constant returns to scale: doubling S will double Q. Finally, we define the inverse demand curve as $P(Q, Z, Y)$.

Therefore in the second-stage under Cournot competition, each firm solves:

$$\max_{q_i} \Pi_{N,m} \equiv P(q_i, q_{-i}, Z_m, Y_m) \cdot q_i - F_N - C(q_i)$$

where F_N is the endogenous sunk cost associated with N firms entering the market. Without loss of generality (we proved a similar result in the Cournot-Sutton context), assume that in equilibrium, quantities will be symmetric such that $q_i = q_j = q^*$ for all i, j. Because N is already "chosen" in the second stage, we can write $q^* \equiv d(Z_m, p) \cdot \frac{S(Y_m)}{N}$. Then, using this, the profits of each firm is given by:

$$\Pi_{N,m} = P(q^*, q^*, Z_m, Y_m) \cdot q^* - F_N - C(q^*) = \left(P_N - \frac{C(q^*)}{q^*}\right) \cdot d(Z_m, P_N) \cdot \frac{S}{N} - F_N$$

Now, for both the average variable cost $\frac{C(q^*)}{q^*}$ and the fixed cost F_N, let's allow them to be additively separable in components, one of them being dependent on the firm only (respectively AVC and F) and the other component being endogenous on the number of firms (respectively b_N and B_N), such that the total profit of an individual firm, facing a market with N total firms, is defined as:

$$\Pi_{N,m} = (P_N - AVC(q^*) - b_N) \cdot d(Z_m, P_N) \cdot \frac{S}{N} - F - B_N$$

In the first stage, using the free entry assumption, we know that if N^* firms enter a market m, it must be that $\Pi_{N^*,m} > 0$ and $\Pi_{N^*+1,m} < 0$. Conversely, we can look at the entry threshold s_N, which is the minimum additional demand, not already covered by existing $N-1$ firms that is required in order for the Nth firm to enter. This threshold is the value of $s_N \equiv S_N/N$ such that:

$$\Pi_{N,m} = 0 \Leftrightarrow \frac{S_N}{N} = \frac{F + B_N}{(P_N - AVC_N - b_N)d_N}$$

We can finally also look at the ratio of successive thresholds:

$$s_{N+1}/s_N = \frac{F + B_{N+1}}{F + B_N} \frac{(P_N - AVC_N - b_N)d_N}{(P_{N+1} - AVC_{N+1} - b_{N+1})d_{N+1}}$$

Thus, in a fully competitive market, a firm would enter each time its cost to enter could be recovered in the market, making s_{N+1}/s_N tend to 1, as $N \to \infty$.

Empirical strategy

The threshold equations derived above ask for a lot of observed variables, namely prices, costs (variable and sunk) and more. In reality, it is very difficult to observe all these at the same time for all firms in a market, thus we need a model that could allow for less information, which is the essence of [**?**].

Consider a reduced-form profit function given by:

$$\Pi_{N,m} = \underbrace{\underbrace{S(Y_m, \lambda)}_{\text{size}} \cdot \underbrace{V_N(Z_m, W_m, \alpha, \beta)}_{\text{variable profit}} - \underbrace{F_N(W_m, \gamma)}_{\text{fixed}}}_{\bar{\Pi}_{N,m}} + \epsilon_m$$

which has an endogenous component $\bar{\Pi}_{N,m}$, and an error-term ϵ_m that is not dependent on N. This should raise some memories to anyone who covered the demand estimation part of the course. In fact, assuming a specification on the error term reveals the model as an ordered probit model, very similar to the logit or multinomial probit models. To see this, consider a market with N firms, this means that:

$$\bar{\Pi}_{N,m} + \epsilon_m > 0 \text{ and } \bar{\Pi}_{N+1,m} + \epsilon_m < 0$$

which is equivalent to:

$$\bar{\Pi}_{N,m} > \epsilon_m > \bar{\Pi}_{N+1,m}$$

Assuming ϵ_m is drawn from an iid normal distribution with mean 0 and variance σ^2, we have that:

$$\Pr\left[N_m\right] = \begin{cases} \Phi(\bar{\Pi}_{N,m}) - \Phi(\bar{\Pi}_{N+1,m}) & \text{if } N_m > 1 \\ 1 - \Phi(\bar{\Pi}_{2,m}) & \text{if } N_m = 1 \end{cases}$$

Before going on to the estimation of the model, we need to look at the reduced-form of each component of the profit function:

- The market size includes a combination of market size characteristics Y_m, such as the population, the neighbouring population, growth, commuting possibilities, etc.
- The variable profit per capita is defined as: $V_N = \alpha_1 + X\beta - \sum_{n=2}^{N} \alpha_n$ where:
  - X is a vector of relevant economic variables such as demand characteristics for the product (Z) and cost-shifters (W).
  - $\alpha_n \geq 0$ is an intercept component such that each firm entering a market has a negative effect on profits.
- The endogenous fixed costs defined as: $F_N = \gamma_1 + \gamma_L w_L + \sum_{n=2}^{N} \gamma_n \cdot \gamma_n$

Finally, the estimation is performed using maximum likelihood. In fact, for each observation of the market, we use the probability function defined above. This gives us a likelihood function as a function of all observed variables described in the list above. Taking the log of it yields the log-likelihood to be maximized in order to estimate the parameters of interest.

5.3.3 Berry (1992)

The previous paper was lacking heterogeneity in the sense that its results applies for the case where all firms have the same costs, same continuation payoffs, etc. A step in the direction of allowing some differentiation was made by [**?**], where fixed costs can vary across firms. Thus, the behavioral model stays identical, but the profit of firm k can now be divided into a common component $\nu_m(N)$, incurred by all firms in the same way, and an idiosyncratic component $\phi_{m,k}$ applying only to firm k. In particular, we have:

$$\Pi_{N,m,k} = \underbrace{X_m\beta - \delta\ln(N) + \rho u_{m,0}}_{\nu_m(N)} + \underbrace{Z_k\alpha + \sqrt{1-\rho^2} \cdot u_{m,k}}_{\phi_{m,k}}$$

Note that in this equation, the combination of the terms $\rho \cdot u_{m,0} + \sqrt{1-\rho^2} \cdot u_{m,k}$ actually represents the error term, denoted $\epsilon_{m,k}$. In this setting, ρ represents the correlation between error terms across firms in a given market. Finally, [?] assumes that:

$$\epsilon_{m,k} \sim N(0, \Sigma)$$

where Σ is a matrix with all off-diagonal terms equal to ρ.

As in the demand estimation case, this error term is known to the players (the firms) but not to the econometrician. We are still in the static, full information game where all firms know everything about all other firms. However this time we are not in an ordered probit setting since the very structure of the problem is endogenous to the number of firms. To see that, recall the error of the previous model did not include any other subscript than m, while this one includes something about k which is intrinsically linked to the number of firms.

Contrary to the previous models, this one can display multiplicity of equilibria. In fact, although the equilibrium number of firms is unique, the exact firms that enter the market can be different. This implies that we lose information on which firms enter a market. For example, the model might predict that in a given market, two firms will ultimately enter, but you could have firms 1 and 2, firms 2 and 3 or firms 1 and 3. In order to control that issue, the author assumes that firms enter in the order of their profitability. Using that assumption, we can simplify the paper and compute the probability of observing N firms in the market as:

$$\begin{aligned}
\Pr\left[n_m = N | Z_m\right] &= \Pr\left[\epsilon_{m,1}, \ldots, \epsilon_{m,K_m} : \sum_{k=1}^{K_m} \mathbb{I}\{\nu_m(N, Z_m) > \phi_{m,k}\} = N\right] \\
&= \underbrace{\int \ldots \int}_{K_m \text{ times}} \mathbb{I}\left\{\sum_{k=1}^{K_m} \mathbb{I}\{\nu_m(N, Z_m) > \phi_{m,k}\} = N\right\} \mathrm{d}F(\epsilon_{m,1}, \ldots, \epsilon_{m,K_m}, \theta)
\end{aligned}$$

which is a multi-dimensional integral (of K_m dimensions) and as we've seen in the previous two chapters, it is hard to compute. In order to circumvent this computational issue, we could use the same method as in BLP: an inner loop simulating the sample moment of the integral for each value of θ, and an outer loop solving for the value of θ that minimizes the square distance between the observed and simulated number of firms. This technique is called the Simulated Nonlinear Least Squares (or SNLS).

Formally, the outer loop finds

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{M} \sum_{m} \left(N_m - \mathrm{E}\left[n_m | Z_m, \theta, \epsilon_{m,1}, \ldots, \epsilon_{m,K_m}\right]\right)^2$$

where the expectation term is approximated by its simulated sample average over random iid draws of $(\epsilon_{m,1}, \ldots, \epsilon_{m,K_m})$. The simulated sample average is defined as:

$$\mathrm{E}\left[n_m | Z_m, \theta, \epsilon_{m,1}, \ldots, \epsilon_{m,K_m}\right] \approx \frac{1}{S} \sum_{s} n_m^{*,s}(\epsilon^s, \theta)$$

$$\text{where } n_m^{*,s}(\epsilon^s, \theta) \equiv \max_{0 \le n \le K_m} \left\{ n : \sum_{k=1}^{K_m} \mathbb{I}\{ v_m(n, Z_m) > \phi_{m,n} | \epsilon^s \} \ge n \right\}$$

or in words, the maximum number of firms n, such that the number of firms entering the market with a positive profit is greater than n.

To recap the full estimation algorithm, consider these steps:

5.4 Two-period models

5.4.1 Stackelberg's model

Two-period models have been introduced for the first time in entry models, but they have been used in several other settings since, for example in models of investment, contracting, etc. Typically, these models rely on:

- A first period where the agents choose a state variable that will determine the nature of the game in the second period.
- A second period where the game is played following the state of the world decided on the first.

The solution concept used to determine the outcome of the game is called "subgame perfection" which means a Nash equilibrium is chosen at each step played in the game. We also call this equilibrium the Subgame Perfect Nash Equilibrium, or SPNE. In the simple case of a two-period game, the SPNE is quite easy to determine by backwards induction:

- You solve for the NE of the second period game for each potential game played following first period choices.
- Assuming the outcomes computed above are realized, you solve for the first period choice that is a NE.

Two-period capacity game

Consider a typical two-period à la Stackelberg game where a firm chooses a level of capital ($K_1 \geq 0$) in the first period, and the second firm chooses its own level ($K_2 \geq 0$) in the second period. A firm that chooses a capital level of 0 is choosing to stay out of the market. A firm that chooses a level strictly greater than 0 enters the market, although firm 2 would have to pay a fixed cost of entry equal to F.

The profits of firm 1 are:

$$\pi_1(K_1, K_2) = K_1 \cdot (1 - K_1 - K_2)$$

while the profits of firm 2 are:

$$\pi_2(K_1, K_2) = \begin{cases} K_2 \cdot (1 - K_1 - K_2) - F & \text{if } K_2 > 0 \\ 0 & \text{if } K_2 = 0 \end{cases}$$

In order to solve the game, we proceed by backwards induction. In the second period, we consider the decision of firm 2 over the level K_2. It will choose a level $K_2 > 0$ if and only if the profit associated is greater than 0, meaning that:

$$K_2(1 - K_1 - K_2) - F \geq 0 \Leftrightarrow K_2(1 - K_1 - K_2) \geq F$$

In that case, firm 2 will choose:

$$K_2^* = \arg\max_{K_2} K_2(1 - K_1 - K_2) - F$$

which is the solution to the following FOC:

$$1 - K_1 - 2K_2^* = 0 \Leftrightarrow K_2^* = \frac{1 - K_1}{2}$$

Plugging this solution into the participation condition above, we get that:

$$\frac{1-K_1}{2}(1-K_1-\frac{1-K_1}{2}) \geq F \Leftrightarrow \frac{(1-K_1)^2}{4} \geq F \Leftrightarrow K_1 \leq 1-2\sqrt{F}$$

and thus the optimal level K_2^* is defined as:

$$K_2^* = \begin{cases} \frac{1-K_1}{2} & \text{if } K_1 \leq 1-2\sqrt{F} \\ 0 & \text{if } K_1 > 1-2\sqrt{F} \end{cases}$$

Now that the second period is solved, we can go into the first period. Firm 1 will choose a level K_1 to maximize its profits, which are now known to be:

$$\pi_1(K_1) = \begin{cases} K_1 \cdot \left(1-K_1-\frac{1-K_1}{2}\right) & \text{if } K_1 \leq 1-2\sqrt{F} \\ K_1 \cdot (1-K_1) & \text{if } K_1 > 1-2\sqrt{F} \end{cases}$$

From this perspective, it is clear that, for a given K_1, allowing the other player to enter the market is worse than deterring entry. However, deterring entry can only be done for $K_1 > 1-2\sqrt{F}$. This creates a discontinuous profit function where deterring entry is always better but can be achieved only at some points that might yield a lower profit than allowing firm 2 to enter. The following graph represents this discrepancy nicely:

5.4.2 Empirical work