Identification and Prices (Cuteonomics Series)

Mallick Hossain and João Granja

Introduction

Identification is the most fundamental aspect of empirical research. However, the intuition is not simple to develop and requires a careful thought process and approach. The most valuable tool to have is the patience to work through simple examples and then carefully think of how to apply the strategy to your specific example. Today's example will walk through the intuition and math behind identifying consumer preferences over product characteristics. We will specifically focus on preferences over price and size.

Model

Most papers/textbooks/lectures will start with a toy model and everything falls out of the model naturally. While that approach is elegant and clearly illustrates how a specific thing is identified, it can be less helpful because it is a retrospective verification study as opposed to a bottom-up development that illustrates the proper thought process. While messier, we think it is more informative to start with one of two approaches: (a) start with a setting and see what can be identified and grow from there or (b) given a goal, identify the ideal experiment to estimate it and then working from the ideal experiment to the real world. We will focus on approach (a) in this lecture. (b) will be saved for a future lecture.

Mallick is really interested in understanding how consumers trade-off a product's price against its size. He had a large amount of shopping data, so it seemed obvious that he could identify this parameter. After all, he could see thousands of households make many choices across a variety of stores over time in a variety of markets across a variety of products. Variation was coming from everywhere! At least one of those had to identify what he needed. However, he had not carefully thought about the source of his identification, so maybe he was right, or maybe not and any identification was just coming from random noise. He could not tell the difference.

Today, we are focusing on fundamentally understanding what can be identified given a particular setting. Mallick's intuition was that looking at a single store and observing consumer choices across different brands and sizes had to buy him something! Let's start there.

First, let's start with a basic representation of a consumer's utility. The consumer (i) cares about the price, size, and brand of a product (j).

$$u_{ij} = \alpha p_j + \beta size_j + \gamma B_j + \epsilon_{ij} \tag{1}$$

This step is crucial, even though it might seem trivial. Writing this down clearly illustrates how many unknowns you are trying to estimate. At the end of the day, an identified model is one that can be solved mathematically, and from math, we know that you can only solve for as many unknowns as you have equations. Hence, we will need at least 3 equations to get our 3 parameters.[1]

[1] Already, we are making an assumption. Either (a) we are assuming that price, size, and brand are continuous so there's only 3 parameters, (b) if brand is discrete, then we will only have 3 parameters if there are 2 brands. If we have B brands, then we have B-1 coefficients on those dummies. Already, we see that assumptions on whether your variables are discrete or continuous can get you into trouble. Be careful!

Setting

The goal of this section is to build, step-by-step, the intuition of exactly what identifies different aspects of the model. This section is also meant to illustrate the value of starting with a simple model. Most students (the authors included!) are seduced by wanting to skip to the most technical, flashy, and complex models. Don't! Start simple, figure out what you need and stop... or keep going.[2] It's up to you.[3]

Attempt 1: 1 Product, 2 Sizes

Let's start with the most simple setting. We care about price and size, so what could we figure out if there was 1 product sold in 2 different sizes at the store? What is the intuition here? We see some customers buying the big size and some buying the small size. There is only one type of variation here. Some buy big and some buy small. We cannot identify 2 parameters off of 1 type of variation. Another way to think about this is by considering competing explanations. Mallick thinks people are highly sensitive to price and less sensitive to size. João thinks the opposite. Simply seeing some customers buy big sizes and some buy small sizes does not give us enough information to figure out which explanation is right.[4]

The takeaway from this is that there is only one source of variation, the choice between large and small. That's not enough to get size and price preferences.

Attempt 2: 2 Products, 2 Sizes

Let's add some more information and see if we can do any better in this case. Instead of 1 product, let's have 2 products: Brand A and Brand B. Each offers 2 sizes. Immediately, we might be worried because by adding a new brand, we have to now worry about a brand preference in addition to our size and price preferences. Do we have enough variation? Let's see what kinds of variation we can deal with. To help, let's look at the illustration:

	Brand A	Brand B
Large	p_{AL}	p_{BL}
Size	p_{AS}	p_{BS}

Each product has a price and a size and we observe customers making choices of various products. What variation do we have?

- For customers choosing brand A, we see some picking large and some picking small. Same for brand B. This is the variation the columns of the table.
- For customers choosing Large, we see some choosing brand A and some choosing brand B. Same for Small sizes. This is the variation in the rows of the table.

Are we stuck? It looks like we only have 2 kinds of variation and we have 3 unknowns. Not yet! Remember that we got stuck in the first case because we could not distinguish between price and size preferences given only 1 brand. We actually haven't fully utilized the fact that we have 2 brands offering the same sizes. In particular, the final piece of variation is the variation in the differences between how brand A customers choose between sizes and how brand B customers choose between sizes. Alternatively, the same variation is

[2] Let's be honest though. If you keep going, you're just showing off at that point. There's an elegance to choosing the simplest model for your needs and no simpler. No more complex either. It also has the added benefit of being accessible to a wider audience.

[3] Or your adviser. Probably your adviser.

[4] In an unrealistic scenario, if prices were the same between the two, then we could credibly identify size preferences, but that's just as a result of shutting down the price channel. Also, there's likely other issues in that it would make no sense for a small and a large package to be priced the same.

captured in looking at the differences between how Large customers pick between brands and comparing that to how Small customers pick between brands.

Here is an illustration. If we see twice as many people choosing Large A compared to Small A, then that suggests that people prefer larger sizes. However, if we see three times as many people choosing Large B compared to Small B, then this suggests there is something else going on. If everything was about size preferences, then within a brand, we would expect to see similar sorting. The difference in sorting must be due to either something different in how the products are priced. That's using the differences between sizes within brands (comparing differences of the columns).

The alternative approach is to look at differences between brands within sizes (comparing differences of the rows). These give you the same answer, so it's a matter of preference and interpretation. In this case, you look at Large customers and see that twice as many are buying Brand A compared to Brand B. This suggests a strong preference for brand A. For Small customers, if three times as many people are buying brand A as brand B, this must be due to something besides brand preferences, namely price.

Solving for the Price Parameter

Let's work through the math and the intuition step by step. First off, we might think that we have 4 equations since there are 4 products.[5]

$$\begin{aligned} S_{AL} &= \frac{\exp(V_{AL})}{\sum_J \exp(V_j)} \qquad S_{AS} = \frac{\exp(V_{AS})}{\sum_J \exp(V_j)} \\ S_{BL} &= \frac{\exp(V_{BL})}{\sum_J \exp(V_j)} \qquad S_{BS} = \frac{\exp(V_{BS})}{\sum_J \exp(V_j)} \end{aligned} \tag{2}$$

where $V_j = \alpha p_j + \beta size_j + \gamma B_j$ is the deterministic part of the utility function.

4 equations and 3 unknowns! Looks like we're over-identified! Actually, we are not. We only have 3 equations. Why? Since we are looking at market shares, once you know 3 of them, you can determine the last one since all the shares have to add up to 1. 3 equations and 3 unknowns. We are still able to solve this system.

All of the above intuition was focused on the difference in market shares, so let's construct those differences. Since we have a lot of exponentials floating around, we will look at the difference in log shares instead to make things much nicer.

We will construct the difference in log shares for each size of Brand A:

$$\begin{aligned} \log(S_{AL}) - \log(S_{AS}) &= V_{AL} - V_{AS} && (3)\\ &= (\alpha p_{AL} + \beta size_{AL} + \gamma B_{AL}) - (\alpha p_{AS} + \beta size_{AS} + \gamma B_{AS}) && (4)\\ &= (\alpha p_{AL} + \beta size_{AL}) - (\alpha p_{AS} + \beta size_{AS}) \quad \text{same brand cancels} && (5)\\ &= \alpha(p_{AL} - p_{AS}) + \beta(size_{AL} - size_{AS}) && (6) \end{aligned}$$

By looking at the shares within a brand, the brand preference cancels out leaving us with only 2 parameters. What's the 2nd equation that will help us solve? We just do this for brand B:

[5] We will assume a basic multinomial logit model where the error is iid Type 1 extreme value. Since the focus of this lecture is to develop intuition and not dig into the details of discrete choice or logit, the only fact you need to work through the math is that the market shares of each product can be represented as follows: $S_j = \frac{exp(\alpha p_j + \beta size_j + \gamma B_j)}{\sum_J exp(\alpha p_j + \beta size_j + \gamma B_j)}$. What's the intuition? We would expect that the market shares should be some function of the utility with items that give higher utility getting higher market shares. The extent that people deviate from picking the highest utility item is due to the ϵ shocks. In a multinomial logit model, the shocks conveniently deliver the equation above. Like we said, even if you have not learned discrete choice, all you need is the market share equation.

$$\log(S_{BL}) - \log(S_{BS}) = V_{BL} - V_{BS} \tag{7}$$
$$= (\alpha p_{BL} + \beta size_{BL} + \gamma B_{BL}) - (\alpha p_{BS} + \beta size_{BS} + \gamma B_{BS}) \tag{8}$$
$$= (\alpha p_{BL} + \beta size_{BL}) - (\alpha p_{BS} + \beta size_{BS}) \quad \text{same brand cancels} \tag{9}$$
$$= \alpha(p_{BL} - p_{BS}) + \beta(size_{BL} - size_{BS}) \tag{10}$$

By squinting just a little, we know that both brands offer the same sizes, so by taking the difference of these two equations, the β term will cancel out leaving us with an equation related the share differences to the price differences which gives us the identification of the price parameter.

$$[\log(S_{AL}) - \log(S_{AS})] - [\log(S_{BL}) - \log(S_{BS})] = \alpha\left[(p_{AL} - p_{AS}) - (p_{BL} - p_{BS})\right] \tag{11}$$

While the above looks a little messy, remember the intuition. The left-hand side has 2 terms. The first is the difference in shares between large and small sizes of brand A. The second is the difference in large and small shares for brand B. If these shares are different (e.g. 2x people prefer large A to small A while 3x people prefer large B to small B), then something must be generating this. It can't be brand because that would not affect the relative shares within a particular brand. It can't be size since we're looking at the same sizes choices across brands. That leaves us only with price. These products must be priced in a certain way that is generating this pattern. That is what the right hand side is saying. This differential sorting must be driven by the difference in the difference of prices between the sizes. Put in more concrete terms, if the large size only cost \$1 more for both brands, then the RHS is 0 and we are stuck. This also fits intuition. If we saw the sorting described above (2x people like Large A compared to Small A versus 3x people preferring Large B to Small B), we would think it has to be that large B is a better deal compared to small B. Maybe large B only costs \$0.50 more which makes it more attractive to customers that like B. If it's cheaper to up-size in B compared to A, that would explain why there is this discrepancy in the size popularity between brands.

Solving for Size Parameter

Now that we have solved for our price parameter α, we can solve for the remaining parameters. While this is just math, we think it is important to develop the intuition of why this works. If we look at the relative shares of the large and small size within a brand, this will give us our size parameter. Think of it in the following way: if we know a person's price sensitivity, then we can predict how they would choose between the large and small product based on the price. If they deviate from that prediction in a systematic way, the only explanation is that they prefer one size over the other. The difference between our predicted shares based only on the price and the actual shares tells us how customers value price.

Solving for Brand Preference

We can also solve for brand preference but by looking at the difference in shares between brands within a particular size. Given prices, we can predict how consumers will choose between large A and large B or small A and small B. To the extent that they deviate systematically from this prediction, that could only be attributed to preferences over the brands.

Attempt 3: X Brands, Y Sizes

We showed that in our 2x2 case, we were exactly identified and we walked through the intuition of what kind of variation is identifying each parameter. What happens if we increase the number of sizes offered? What about if we increase the number of brands? As is our theme, let's think very carefully about what are the good and bad things when we do this. At the core, we need to be highly aware of how many parameters we are estimating and how many equations we have.

Starting with equations, in a general sense, how many equations will we have if we have X brands and Y sizes? Using the same logic as above, we will have $X * Y$ market share equations. However, since they must sum to 1, we only have $X * Y - 1$ equations that can be used to solve for our parameters. In our 2x2 case, that meant we had 3 equations to use. Once we have our equations, the solution is algebra. We used the differences above to better highlight the intuition behind each parameter.

What about parameters? This is where we have to be careful and this is where deciding between discrete or continuous variables can get you into trouble. Let's start by looking at price. It makes sense that price is a continuous variable. Therefore, no matter how many products we have, each one has some price which is continuous. Therefore, there is only one price parameter to estimate, no matter whether we grow the number of products or the number of firms.[6]

What about size? If we are modeling it as a continuous variable, like weight or volume, then we only have 1 parameter no matter the number of products. However, if sizes are discretized, like we had in our example, then you will have $Y - 1$ different parameters to estimate. The same holds true for brand, though you would be hard-pressed to argue that this should be continuous. Brand makes most sense as a discrete variable. Therefore, you will have $X - 1$ different brand parameters to estimate.

Cases

Let's do the accounting for our 2x2 exercise and then expand it to see how things grow as we grow different dimensions of the space.

- *Continous price, continous size, discrete brand*
  - 3 equations (2 * 2 - 1 = 3)
  - 3 parameters: 1 price parameter, 1 size parameter, 1 brand parameter
  - Exactly identified
- *Continous price, discrete size, discrete brand*
  - 3 equations (2 * 2 - 1 = 3)
  - 3 parameters: 1 price parameter, 1 size parameter, 1 brand parameter
  - Exactly identified

Let's look at 3 sizes

- *Continous price, continous size, discrete brand*
  - 5 equations (2 * 3 - 1 = 5)
  - 3 parameters: 1 price parameter, 1 size parameter, 1 brand parameter
  - Over-identified
- *Continous price, discrete size, discrete brand*
  - 5 equations (2 * 3 - 1 = 5)
  - 4 parameters: 1 price parameter, 2 size parameters, 1 brand parameter
  - Over-identified

Y sizes

- *Continous price, continous size, discrete brand*
  - $2Y - 1$ equations
  - 3 parameters: 1 price parameter, 1 size parameter, 1 brand parameter
  - Over-identified as long as $Y > 2$

[6]This actually leads to an interesting kernel of truth. If you think that people are only making a decision based on one continuous characteristic, like price, then the minimum amount of information you need is people making a choice between 2 products. That's it. Honestly, you should never do this, because then you end up literally comparing apples and oranges only based on price. Why do people buy apples? They are cheaper! In the real world, there are a lot of other factors to consider as you are building your model. This is the "art" part of economics. The math tells you that you can estimate price sensitivity only based on observing choices between 2 products. However, experience tells you that people choose products based on a wide variety of characteristics and it is up to you as the economist to decide what are the important factors. If you think there is some underlying preference for apples relative to oranges, then as we outlined in Attempt 1, you will need more variation to estimate that.

- *Continous price, discrete size, discrete brand*
  - $2Y - 1$ equations
  - $Y + 1$ parameters: 1 price parameter, Y-1 size parameters, 1 brand parameter
  - Over identified as long as $Y > 2$

Just looking at sizes illustrates an important fact: continuous variables do not scale with the space, only discrete variables do.

What if we have 3 brands and Y sizes?

- *Continous price, continous size, discrete brand*
  - $3Y - 1$ equations
  - 4 parameters: 1 price parameter, 1 size parameter, 2 brand parameters
  - Over-identified as long as $Y > 1$
- *Continous price, discrete size, discrete brand*
  - $3Y - 1$ equations
  - $Y + 2$ parameters: 1 price parameter, Y-1 size parameters, 2 brand parameters
  - Over identified as long as $Y > 3$. Exactly identified if Y = 3.

What about X brands and Y sizes?

- *Continous price, continous size, discrete brand*
  - $X * Y - 1$ equations
  - $X + 1$ parameters: 1 price parameter, 1 size parameter, X - 1 brand parameters

X (brands)	Y (size)	Identification
$\forall X$	1	Not identified
1	2	Not identified
1	3	Exactly identified
2	2	Exactly identified
≥ 2	> 2	Over-identified

- *Continous price, discrete size, discrete brand*
  - $X * Y - 1$ equations
  - $X + Y - 1$ parameters: 1 price parameter, Y-1 size parameters, X-1 brand parameters
  - If $X = 1$ or $Y = 1$, not identified
  - If $X = 2$ and $Y = 2$, exactly identified
  - If $X \geq 2$ and $Y \geq 2$, over-identified

Summary

We have walked through how to carefully think about identification in a discrete choice setting. While the intuition will take much practice and patience to fully internalize, we have tried to lay the groundwork for how to approach these kinds of problems. On one level, it is just an accounting exercise keeping track of the number of equations and parameters. However, that over-simplifies the problem because the obstacles and mis-steps can be subtle. Keeping track of equations is relatively straightforward because it is the number of products minus 1 to allow for the fact that all shares have to sum to 1. However, the parameter space is less obvious especially because while continuous variables do not scale with the product space, discrete variables can and you have to be vigilant in recognizing that.

What are the caveats here? There are many. First, we have primarily presented the identification arguments and intuition, but there is a whole other dimension related to whether this is an economically meaningful model. In this particular scenario, we can imagine that one clear omission from the utility function is that people also care about how much "stuff" they are buying, especially in this setting. People might like

large quantities because it means they have more "stuff" to consume. We hope that we have given you an intuition on how to think about these additional dimensions because they will come up when you take off your identification hat and put on your economist hat.

Another important caveat is that there are many other kinds of variation that we are clearly not utilizing. For example, you might have consumer-level choice data. As we have illustrated, none of that variation is necessary because you basically aggregate all of that data to market shares and solve for your parameters. That seems like an awful waste of valuable information. By our model's construction, that data is not useful because we are assuming that all customers have the same preferences and any variation only comes through the error term, so there cannot be any other kind of systematic variation in consumers. This explicitly excludes any kind of useful information that we could extract from demographics, income, or even repeated purchases by the same customers. This also means any variation across markets can only be due to noise.

In the next lecture, we will approach the question of how incorporate other kinds of data and what kinds of parameters we can identify as we try to build out a more realistic model of consumer behavior.