Identification Pitfalls (Cuteonomics Series)

Mallick Hossain and João Granja

Introduction

In the last lecture, we discussed the intuition for how to identify preference parameters of a utility function. We briefly touched on the importance of stepping back from your identified model to see if the model actually makes sense or if there is something important that is missing. We highlighted one important omission from Mallick's price/size/brand utility function and we will spend this lecture exploring the underlying assumptions and insights.

Missing "Stuff"

In the last lecture, we were estimating the following utility function:

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

This function assumes that the only characteristics driving choice are a product's price, size, and brand. Everything else is noise. However, this immediately leaves out probably the most important reason why people buy big packages, namely, you get more "stuff" to consume. People buy gallons of milk instead of 8oz bottles because they prefer more milk to less, not because they like a big jug instead of a tiny bottle. Our utility function does not capture that fact, and, while well identified, probably not a good model. We can do better.

Before we correct for this omission, let's think carefully about what this would mean for our current model.[1]

To give us a starting point, let's make the (seemingly obvious) assumption that people prefer more stuff to less. How would this manifest itself in our model?[2] More stuff likely comes in bigger packages and likely costs more. Therefore, if people are buying bigger or more expensive things because they have more "stuff", then we are likely thinking people's preference for price or size is more positive than it should be.[3] In some extreme cases, this preference might be so strong as to actually manifest itself as a positive preference for price or size. Then we could be making an unrealistic conclusion that households actually prefer expensive items or humongous packages.

If you care about accurately estimating price sensitivity or size preference, then this is bad news and you will want to correct it.[4] How can this be appropriately corrected? The easiest way is to directly control for the actual quantity of the product. Let's rewrite our utility function:

[1] Arguably, the following paragraph is probably one of the most important skills of an economist and for you on the job market. Honestly, everyone knows your model is not perfect and probably pretty bad. You probably know your model is not perfect and pretty bad. However, it is good for something. You need to know what that something is. Also, people are going to ask why your model doesn't reflect the real world, often by giving some personal story. In most cases (assuming you've done some decently careful work on your own model), they are not asking you to add in their particular feature into your model. And you probably should not if your object of interest is something very different. However, what they are asking is for you to have carefully thought about how that omission could affect your model. Will it weaken your results? If so, then assuming you got something significant, then you have still demonstrated your story has value. Will it bias your results upward? Then you might be getting significance where there is none. You would want to think carefully about how strong this bias might be, or maybe this is a case where you should directly incorporate this feature into your model to remove the risk of identifying something that is not there. This is an extremely valuable skill to have and not one explicitly taught in graduate school.

[2] Yes, the following intuition is just a discussion of omitted variable bias. You will never escape it.

[3] To the extent that both of these coefficients are likely negative, this would be downward biasing our results, because it is making the absolute magnitude smaller. However, it is pushing our results upward from a more negative number. Remember a smaller negative is a bigger number.

[4] This is exactly Mallick's situation.

$$u_{ij} = \alpha p_j + \beta size_j + \gamma B_j + \delta q_j + \epsilon_{ij} \tag{2}$$

Great! Now we are directly letting utility be a function of quantity as well as price and size. Taking a quick tally of equations and parameters, if we only have 2 brands and 2 discrete sizes, then we have a grand total of 4 parameters to estimate! Unfortunately, 2 brands and 2 sizes only gives us 3 equations. We would need at least 2 brands and 3 sizes to be exactly identified and more sizes would make us be over-identified. Given typical product assortments, that is a quite realistic dataset to have.

Is it as simple as that? Well, almost. Once again, this all comes back to econometric fundamentals. Mathematically, it seems like this should all work out cleanly and identification is assured. We had an omitted variable, we added it to our regression, so we have fixed one problem. We checked that we have identification. What could we be missing? In most scenarios, we would be okay, but in this particular one, we have made a mistake. In particular, we have to check for multicollinearity. That would kill our identification no matter how much data we had. In particular, we should be worried about the correlation between a package size and its quantity. In many instances, these will be highly correlated.[5]

For a wide range of products, the strength of this correlation would kill any hope for disentangling size preferences from quantity preferences. And the importance of quantity preferences means that you would have to include it in the utility function. This may be a rare case in that most discrete choice models can safely ignore quantities if consumers are choosing 1 out of an assortment of products where quantity does not matter (i.e. cameras, cell phones, etc.). In these cases, quantity is essentially controlled for because there is no quantity difference between the products (e.g. Nikon doesn't have a different quantity of "camera" than a Canon). The other cases basically assume that size is not an important feature and can be safely ignored (e.g. size is not the reason why a household buys 1 yogurt instead of 5 yogurts). Regardless of if you think this is an important feature to model, it illustrates an important part of model development and the potential pitfalls you must be aware of as an economist.

Fixing "Stuff"

How do we solve this impasse? We need to include both quantity and physical size, but we risk identification problems from multicollinearity. One possibility is that you stop here and find a different research question.[6] The other possibility is to think carefully about how to separate these two features. We need something that will effectively separate the physical size of a product from the actual quantity inside. Thinking carefully, there are a couple of possibilities. First, are there instances where the same package has different amounts of "stuff"? Actually, yes! And there are 2 different ways to get at it. The first possibility is realize that the same package might be very different for different households. For example, a 20-pound bag of rice is the same size package to a single household compared to a family of 4. However, these are very different quantities if we measure quantity by how long the package will last, which is arguably the more relevant measure for various food products. This provides the necessary variation to identify size preferences. The intuition is as follows: we compare the shares of the same product bought by big households and by small households. This is the same product, so the brand, size, and price are all the same. The only difference that could generate different purchasing between big and small households is that the product will last for different times for each household. It is important to note that this identification rests on the assumption that big and small households have the same preferences over size. If there are differences in their size preferences, all of this breaks down.

[5] Actually, using some basic physics/chemistry, you can construct the exact relationship. An object's *density* is defined as its mass per unit volume ($\rho = \frac{m}{V}$). Density is an unchanging physical property of an object, so the relationship between the quantity of an object, or its mass is just the density times the volume. Put in our economic notation, $q = \rho * V$. Here we actually have perfect multicollinearity. If we were buying blocks of gold, this would be the case. For other products, there is often some kind of packaging that doesn't scale precisely with the quantity, but it would be close enough to generate a highly correlated relationship.

[6] This is an important skill to have. While painful, you should try to kill your ideas as quickly as possible. If it survives, then it's probably a good one to research.

Can we identify the size preference without resorting to consumer-level data? Yes! The secret is to find products that break this high correlation. For any kind of food product, this is immensely difficult, but if you search more widely, you can find good candidates. The cleaning aisle of your favorite store is probably the easiest place to start. Why? These products often break the size/quantity link by offering various concentrations of their product.[7] Given the wide range of concentrations available, there can be different "quantities" crammed into the same 32oz container. This breaks strong multicollinearity and recovers identification.[8] However, this identification strategy requires that prices be nonlinear in quantities. If prices are linear, then you have exactly the same multicollinearity problem as before, except between prices and quantities instead of between quantities and sizes. The good news is that this happens quite often in a variety of settings because these kinds of products tend to exhibit bulk discounts which induce exactly the nonlinear price schedule needed to avoid multicollinearity.

Summary

In this lecture, we have outlined the other checks that you must do as a researcher to ensure that your identification strategy will work. Even after ensuring your model is "theoretically" identified by accounting for equations and unknowns, you must put on your economist hat and your econometrician hat. As an economist, you must check that your model is capturing the kinds of variation that you want to explain and understand and that you are not omitting major features that will bias your results. As an econometrician, you should be careful about how your variables are related to avoid potential multicollinearity. One looming risk that we did not discuss is endogeneity. This will be covered in a future lecture.

[7] This breaks the "density" equation that I outlined in an earlier footnote because more concentrated soaps pack more "cleaning power" per milliliter. For foods, it's near-impossible to pack more "apple" into an apple (i.e. cramming 2 pounds of apple into 1 pound is impossible.). Different cuts of meat kind of break this, but then you have to deal with how the different cuts are of varying quality or require different amounts of work. For example, chicken thighs and boneless chicken thighs may weigh the same but have different quantities of meat. Boneless thighs take less work to prepare and cook though. This should still be identified though.

[8] Mallick uses this variation in toilet paper to separately identify the storage/transportation costs from quantities.