Likelihood Methods II: Multiple Discrete Choices

Paul Goldsmith-Pinkham

February 16, 2023

Today's topic: multiple discrete choice

- We'll now examine multiple discrete choice problems
- Much of this discussion is very IO-adjacent
  - However, many of these ideas are important for non-IO problems, e.g. multiple IVs and Roy models
  - Moreover, these tools are very promising in fields that have not yet used them
- Issues with choice problems that we'll discuss:
  - Independence of Irrelevant Alternatives (IIA)
  - Choice sets and consideration sets
  - Inconsistency of fixed effects and its consequences

Thinking about multiple choices

- Consider the following problem: we observe choices for individuals $Y_i = j$, $j \in \Omega = \{0, 1, \ldots, J\}$, where $J + 1 = |\Omega|$ is the total number of choices.
  - Importantly, the order of the choices has no particular meaning. This could be red bus, blue bus and car as transportation choices.

- We observe three types of characteristics:
  1. X_i (individual chaaracteristics, invariant to choices),
  2. X_j (choice characteristics)
  3. X_{ij} includes individual-by-choice characteristics

- Can write X_i as X_{ij} by interacting with choice fixed effects
  - Note that when $J = 1$, we collapse down to binary choice

Modeling multiple choices

- Recall from last class that there are two ways to think about how we think about the discrete choice problem. These are not mutually exclusive.

- The first is a statistical view. How do we model the classification of a particular choice.
  - In the binary choice problem, there is only one parameter that needs be known, conditional on X_i: $\pi(X_i) = Pr(Y_i = 1|X_i)$
  - With more than two choices, the dimensionality becomes more complicated. We now have $\pi_j(\mathbf{X}), j = 2, 3$ for 3 choices.

- How should we parameterize how other choices' characteristics affect each other?

- Most of the models we will discuss today will make very specific restrictions on how choices affect one another
  - These are not innocuous choices, as we'll see.

The naive approach

- If we want to estimate simple treatment effects, we could focus on binary outcomes
- For exmaple: we have a randomly assigned treatment T, and J choices. What is the effect of T on $Pr(Y_i = j)$?
  - $\tau_j = Pr(Y_i = j|T_i = 1) - Pr(Y_i = j|T_i = 0)$
- There's less information about the substitution patterns of individuals in this form
- Of course, it is still very helpful! And useful when faced with a lot of choices to focus on the effect on one margin.
- However, need more structure to estimate relative choice substitution across outcomes
  - E.g. what is the effect of T on choosing j conditional on choosing j or k

What's the estimand/counterfactual?

- What counterfactual question are we interested in?
  1. How does changing X_{ij} affect the probability of choosing choice j relative to all other choices?
  2. How does changing X_{ij} affect the probability of choosing choice j *relative* to choice k?
  3. How does adding or subtracting one of the choices (with difference in X_{ij}) change the $J + 1$ choice probabilities?

- An important question is under what settings are these questions identified. In the examples we'll look at, there are answers that fall out (at least for 1 and 2) but they may be too driven by the parametric assumptions.
  - See Berry and Haile (2016) for a discussion of identification in product markets in non-parametric settings.
  - They show that there are two specific conditions that need to hold in the structure of the problem, but allow for very general structure in the distribution of the shocks.

Modeling multiple choices

- A second way to view this is as an structural (economic) choice problem (pioneered by McFadden). Consider a set of utilities U_{ij} (unobserved) such that the

$$Y_i = \arg\max_{j \in \Omega} U_{ij}$$

- E.g., person i chooses j if it's the choice that maximizes the utility amongst all $J + 1$ choices.
  - Note the similarity to the Y_i^* in the binary case

- If we make the assumptions:
  1. $U_{ij} = X_{ij}'\beta + \epsilon_{ij}$
  2. ϵ_{ij} are independent across choices and individuals, and distributed Type-I extreme value

  then we get the McFadden conditional logit model:

$$Pr(Y_i = j | X_{ij}) = \frac{\exp(X_{ij}\beta)}{\sum_k \exp(X_{ik}\beta)}.$$

The impact of price

- In many choice problems, a key parameter we're interested in is a price elasticity
  - A key variable in X_{ij} is p_j
  - This term can be something else, but price matters quite a bit in IO

$$Pr(Y_i = j|X_{ij}) = \frac{\exp(p_j\gamma + X_{ij}\beta)}{\sum_k \exp(p_k\gamma + X_{ik}\beta)}.$$

- Own price elasticity is easily constructed from this:

$$\epsilon_j = \frac{\partial Pr(Y_i = j|X_{ij})}{\partial p_j}\frac{p_j}{Pr(Y_i = j|X_{ij})}.$$

- This is not much different than calculating an average effect. What is more meaningful is that we can think about *cross*-elasticities:

$$\epsilon_{jk} = \frac{\partial Pr(Y_i = j|X_{ij})}{\partial p_k}\frac{p_k}{Pr(Y_i = j|X_{ij})}$$

Independence of Irrelevant Alternatives (IIA)

- A key issue with this formulation of the conditional logit model – the cross-elasticities are identical
  - In other words, $\epsilon_{jk} = \epsilon_{lk}$
  - The effect of shifting price of a different good causes an identical proportionate shift in all choices' market share

- Solve:

$$\epsilon_{jk} = \underbrace{-\gamma Pr(Y_i = j)Pr(Y_i = k)}_{\frac{\partial Pr(Y_i=j)}{\partial p_k}} \times \frac{p_k}{Pr(Y_i = j|X_{ij})}$$
$$= -\gamma Pr(Y_i = k)p_k$$

- Note that this is not a function of j, and hence identical for all other products

Independence of Irrelevant Alternatives (IIA)

- Another way to see this problem: consider the probability of choosing j, conditional on choosing just j and k:

$$Pr(Y_i = j | Y_i \in \{j, k\}) = \frac{\exp(X_{ij}\beta)}{\exp(X_{ij}\beta) + \exp(X_{ik}\beta)}$$

- Note that none of the other choices show up in this probability choice – irrespective of how "similar" the other choices are to j or k.
  - In other words, if a characteristic of the other products changes, the relative share between j and k will stay same

- The canonical example of this is the "car, red bus and blue bus" example.
  - Presumably a person is purely indifferent between red and blue busses.
  - Hence, a shift in the red bus price would cause a bigger substitution from the blue bus than from car users.
  - Conditional logit (in this form) will not account for this

How can we deal with this?

- Better substitution patterns

- Note that this is an economics problem – e.g. we have economic intuition about the market substitution patterns, and we don't think identical cross-elasticities makes sense

- It's also a statistical problem – there is a very strong statistical functional form we have assummed, which was analytically convenient but has somewhat perverse properties

- Will talk about two ways to solve this (there are more in the IO literature):
  1. Nested Logit and Correlated Multivariate Probit
  2. Random Coefficients Logit

Nested Logit and Correlated Multivariate Probit

- One part of the problem comes from the independence of the ϵ across choices
  - Recall that these ϵ effectively rationalize seeing non-zero choices in both directions, conditional on charcteristics

- Recall the blue and red bus case:
  - Getting two independent ϵ draws for the busses is not an intuitive view of bus demand
  - Instead, the blue and bus likely have highly correlated epsilon draws (if not identical)
  - The issue, of course, is what the correlation is within sets

- With the nested Logit approach, you can specify sets (as the researcher), and allow data-driven measures of correlation of the ϵ within these sets.

- The key is that the errors are uncorrelated across choice sets, which preserves the simple logit structure (see Goldberg (1995) for an example application)

Multivarite Probit

- A more general approach is to allow the covariance matrix of the error terms to be flexibly estimated by the data using a multivariate normal
  - E.g. $\epsilon_i = (\epsilon_{i0}, \epsilon_{i1}, \ldots, \epsilon_{iJ}) \sim \mathcal{N}(0, \Sigma)$
  - Directly estimate Σ

- This problem gets hard with many choices (parameter space grows at rate J^2)

- Importantly, do need to normalize one of the variance terms, since the variance matrix is only identified up to scale of one of the terms.

- See McCulloch, Pelson and Rossi (2000) for details in the Bayesian setting, and Train (2009) for simulation discussions in the frequentist case
  - See Hull (2020) for a nice application

Better substitution patterns - Random Coefficients

- Rather than directly target the distribution of the ϵ_{ij}, an alternative approach is to add more richness to the coefficients themselves
  - By adding more random variation in this, it effectively creates a richer substitution pattern

- Now consider a slight extension of our previous model, with β_i varying by individual (in an unobserved way):

$$U_{ij} = X_{ij}\beta_i + \epsilon_{ij}$$
$$U_{ij} = X_{ij}\overline{\beta} + \nu_{ij}, \qquad \nu_{ij} = \epsilon_{ij} + X_{ij}(\beta_i - \overline{\beta})$$

- There are a number of ways to estimate this approach, but notice the key point – subtitution patterns are more richly modeled (and allowed) due to ν_{ij} varying by X_{ij}
  - See McFadden and Train (2000) for details

Symmetric unobserved product differentiation

- The unobservable ϵ is an unobserved valuation of some product characteristic
  - Most models (including the ones we've looked at) have symmetric unobserved product differentiation (SUPD) [Ackerberg and Rysman (2005)]

- Consider our bus and car example – the issue is that adding another bus product in this space should "crowd" the original bus market share
  - E.g. the choices are highly correlated

- This is beyond just IIA's effect of cross-price elasticities – this matters when considering counterfactuals where you add new choices
  - Let X_j define the *characteristic space*
  - If a new product is added in the characteristic space, we think that they should crowd one another. With logit errors, they do not

- Ackerberg and Rysman (2005) propose a solution that incorporates the number of choices directly

- The symmetry of our errors also plays an important role in making the cross price elasticities identical – e.g. $\epsilon_{ij} = \epsilon_{ji}$

Choice sets and consideration sets

- In these discussions, we've assumed that all individual use the same choices
- Reasons why this could not be true are many: attention, knowledge, opportunity
  - Call the subset of choices a consumer focuses on the *consideration* set
  - These can be known (observed) or usually, unknown
- If we assert consideration sets are the full choice set for all individuals, when we see individuals choose certain goods, we view this as reflecting their preferences
  - Or in other words, the counterfactual we generate from this model would imply a certain response
  - E.g. if I never considered going to Harvard in my choice set, a change in its price will be irrelevant for me
- If changes in characteristics affect your consideration set, this can have important implications for counterfactuals

Choice sets and consideration sets

- The general way to view consideration sets in choice problems is

$$s_j(p) = \sum_{C \in P(j)} \pi_C(p) s_j^*(p|C),$$

  where
  - $P(j)$ is the set of consideration sets that include choice j,
  - s_j is the overall choice set of j given prices p,
  - π_C is the probability of consideration set C
  - s_j^* is the choice of j within the choice C

- Note that the key feature of this model is that it can break the symmetry of choice elasticities
  - There is symmetry *within* the consideration set

- Under certain modeling assumptions, it is possible to identify the probabilities of the consideration set choice (Abaluck and Adams-Prassl (2020))

Bias from ignoring consideration sets

Our identification result builds on the insight that imperfect consideration breaks symmetry between cross-price responses (or more generally, cross-characteristic responses). For example, in a model with a default, symmetry would ordinarily require that switching decisions be equally responsive to an increase in the price of the default good by \$100 or a decrease in the price of all rival goods by \$100. Suppose instead that consumers are inattentive and choose the default option unless that good becomes sufficiently unsuitable. Now, switching decisions will be unresponsive to changes in the price of rival goods but more responsive to changes in the price of the default to the degree that these changes perturb attention (Moshkin and Shachar 2002). While the link

...

(2002), this approach has not yet been developed in the generality we consider.[2] Our framework implies that ad hoc attempts to model consideration sets such as fixed effects in utility for products on different shelves or interactions between prices and such fixed effects can still yield misspecified models because they do not relax the symmetry assumption.

Choice sets and consideration sets

- Simple case considered in paper is when there is a base default that people focus on (and ignore other choices)
  - Default Specific Model
- More rich setting: Alternative Specific Choice model
  - Put structure on how a choice is selected into a consideration set
- In both cases, can identify the consideration choice probabilities using price elasticities
- This can be a very important thing to model if your counterfactual relates to changes in the consideration set
  - However, it may not be first order to your problem at hand

BLP

- Another important set of models is known as BLP (Berry Levinsohn Pakes)

$$U_{ijm} = \mu_{ijm} + \delta_{jm} + \epsilon_{ijm}$$

- This exploits knowledge of choices (either aggregated or disaggreated) across many markets

- Can use this knowledge to allow for a lot more market-product specific fixed effects (ξ_{jm}), which gives a richer substitution pattern

- Under distributional assumptions for ϵ_{ijm},

$$s_{jm}(\delta_m, \beta) = \int \frac{\exp(\delta_{jm} + \mu_{ijm})}{\sum_{k \in J_m} \exp(\delta_{km} + \mu_{ikm})} f(\boldsymbol{\mu}|\beta) d\boldsymbol{\mu}_{im} \tag{1}$$

BLP

$$s_{jm}(\delta_m, \beta) = \int \frac{\exp(\delta_{jm} + \mu_{ijm})}{\sum_{k \in J_m} \exp(\delta_{km} + \mu_{ikm})} f(\mu|\beta) d\mu_{im}$$
$$\delta_{jm} = x_{jm}\beta_2 - \alpha p_{jt} + \xi_{jm}$$

- Key insight in BLP is to use a fixed point algorithm to match the estimated market shares in a market, $\hat{s}_m(\delta_m, \beta)$, to the observed market shares
  - This is done iteratively, mapping the shares into a linear index for δ
  - Conlon and Gortmaker (2020) highlight that this can have estimation issues due to convergence in this process
  - Provide Python package to solve this!

Inconsistency in binary choice models

- One issue that arises in many non-linear binary choice model is that many features do not carry over from linear models
  - E.g. interpreting coefficients is more challenging
  - A bigger issue comes from inconsistency of fixed effects
- Consider estimating a panel fixed effects model with binary choice:

$$Y_{it} = \alpha_i + X_{it}\beta + \epsilon_{it}$$
$$Y_{it} = F(\alpha_i + X_{it}\beta)$$

  where we are interested in the parameter β. If we have a short panel (e.g. few time periods), we cannot consistently estimate α_i. However, in the linear case, this does not affect estimation of β
- Unique result (Chamberlain (1987,2010)): for binary outcome case, the only model that consistently estimates β is a conditional logit
- More generally if you have inconsistent fixed effects in your non-linear models, this can cause serious issues (except in special cases like this one)
  - OLS is good!

Underlying structure of discrete choice is valuable in IV settings

- Much of this discussion centered on IO style applications
- But this discussion shows up when thinking about Roy style models
- When we discuss instruments and individuals' choice to take up a policy or not, if the policy is multi-dimensional, this types of models play a huge role
- Recall our discussion of propensity scores for treatment effects
  - If individuals choose between multiple treatment options, this maps directly into a discrete choice setting like what we've discussed today
- Thinking carefully about the counterfactual pattern across will give guidance in more complicated IV settings

Arbitraging IO methods in other settings

- Many fields have discrete choice applications but have not adopted the tools
- The cutting edge of IO tools is quite complex, but this type of structure is very valuable when thinking about complicated choice patterns
- Worthwhile to try to arbitrage these methods in fields that are less exposed to them (e.g. Koijen and Yogo (2019))

Koijen and Yogo (2019)

- Influential paper by Koijen and Yogo (2019) estimates a demand system for financial assets

- The framework is used to study three things:
  1. Distribution of price elasticities with respect to demand shocks (residual demand)
  2. Decomposing variation in asset returns

A Demand System Approach to Asset Pricing

Ralph S. J. Koijen

University of Chicago, National Bureau of Economic Research, and Center for Economic and Policy Research

Motohiro Yogo

Princeton University and National Bureau of Economic Research

We develop an asset pricing model with flexible heterogeneity in asset demand across investors, designed to match institutional and household holdings. A portfolio choice model implies characteristics-based demand when returns have a factor structure and expected returns and factor loadings depend on the assets' own characteristics. We propose an instrumental variables estimator for the characteristics-based demand system to address the endogeneity of demand and asset prices. Using US stock market data, we illustrate how the model could be used to understand the role of institutions in asset market movements, volatility, and predictability.

Koijen and Yogo (2019)

- Influential paper by Koijen and Yogo (2019) estimates a demand system for financial assets
- The framework is used to study three things:
  1. Distribution of price elasticities with respect to demand shocks (residual demand)
  2. Decomposing variation in asset returns

COROLLARY 1. A restricted version of the optimal portfolio (8) under assumption 1 is characteristics-based demand:

$$\frac{w_{i,t}(n)}{w_{i,t}(0)} = \delta_{i,t}(n) = \exp\left\{\beta_{0,i,t}\text{me}_i(n) + \sum_{k=1}^{K-1}\beta_{k,i,t}x_{k,i}(n) + \beta_{K,i,t}\right\}\epsilon_{i,t}(n). \tag{10}$$

We refer to equation (10) as characteristics-based demand because the portfolio weights depend on log market equity, other observed characteristics, and unobserved characteristics. An important question is whether

Koijen and Yogo (2019)

- Influential paper by Koijen and Yogo (2019) estimates a demand system for financial assets
- The framework is used to study three things:
  1. Distribution of price elasticities with respect to demand shocks (residual demand)
  2. Decomposing variation in asset returns

Characteristics-based demand easily captures an index fund. If $\beta_{0,i,t} = 1$, $\beta_{k,i,t} = 0$ for $k = 1, \ldots, K-1$, and $\epsilon_{i,t}(n) = 1$ for all assets $n \in \mathcal{N}_{i,t}$, equation (11) simplifies to

$$w_{i,t}(n) = \frac{\mathrm{ME}_t(n)}{\exp\{-\beta_{K,i,t}\} + \sum_{m \in \mathcal{N}_{i,t}} \mathrm{ME}_t(m)}. \tag{13}$$

This investor is an index fund whose portfolio weights are proportional to market equity, and the intercept $\beta_{K,i,t}$ determines the weight on the outside asset (e.g., cash).

Koijen and Yogo (2019)

- Influential paper by Koijen and Yogo (2019) estimates a demand system for financial assets
- The framework is used to study three things:
  1. Distribution of price elasticities with respect to demand shocks (residual demand)
  2. Decomposing variation in asset returns

2. Investment Mandates and the Wealth Distribution

Let $\mathbb{I}_i(n)$ be an indicator function that is equal to one if asset n is in investor i's investment universe (i.e., $n \in \mathcal{N}_i$). We can trivially rewrite equation (10) for any asset as

$$\frac{w_i(n)}{w_i(0)} = \begin{cases} \mathbb{I}_i(n) \exp\left\{\beta_{0,i}\text{me}(n) + \sum_{k=1}^{K-1}\beta_{k,i}x_k(n) + \beta_{K,i}\right\}\epsilon_i(n) & \text{if } n \in \mathcal{N}_i \\ \mathbb{I}_i(n) = 0 & \text{if } n \notin \mathcal{N}_i. \end{cases}$$

This notation emphasizes that an investor does not hold an asset for two possible reasons. The first reason is that the investor is not allowed to hold the asset because it is not in its investment universe (i.e., $\mathbb{I}_i(n) = 0$). For example, an index fund cannot hold assets that are outside the index. The second reason is that the investor chooses not to hold an asset even though it could (i.e., $\epsilon_i(n) = 0$). For example, an index fund may choose not to hold an asset in the index that is perceived to be overvalued. Thus, $\mathbb{I}_i(n)$ is exogenous under the maintained assumption that the investment universe is exogenous, while $\epsilon_i(n)$ is endogenous through the portfolio choice problem.