Discrete Chocice Models: Individual Data

Chris Conlon

Spring 2023

NYU Stern: Applied Econometrics

Readings

After class please read:

- Ken Train https://eml.berkeley.edu/books/choice2.html
  - Chapters 1-6

Lots of Discrete Choices (in IO and elsewhere)

- Choosing a brand from the supermarket
- Choosing a hospital
- Choosing a school/major
- Choosing where to live

A Purely Statistical Setup: Utility

An individual i has utility for choice j:

$$u_{ij} = V_{ij}(\theta) + \varepsilon_{ij}$$

Idea:

- Make some assumption on $f(\varepsilon_{ij})$
- Estimate $V_{ij}(\theta)$

A Purely Statistical Setup: Choice Probabilities

Consider the choice probability

$$s_{ij}(\theta) = \mathbb{P}(u_{ij} > u_{ik}; \forall k \neq j)$$

Idea:

- Only choose j if it is preferred to all options k.
- Helpful to assume that $f(\varepsilon_{ij})$ is continuous and full support so that ties are measure zero.

A Purely Statistical Setup: Data

Assume that we see individual i makes a:

- Mutually exclusive and exhaustive discrete choice.
- $y_{ij} \in \{0, 1\}$ and $\sum_k y_{ik} = 1$

And that we observe y_{ij} for $i = 1, \ldots, N$ and all j. We can write the likelihood of the dataset as

$$L(\theta) = \prod_i \prod_j s_{ij}^{y_{ij}}(\theta)$$

$$\ell(\theta) = \sum_i \sum_j y_{ij} \log s_{ij}(\theta)$$

Since the cases where $y_{ij} = 0$ do not contribute to $\ell(\theta)$ some people abuse notation and drop y_{ij} or $\sum_j$ and write $\ell(\theta) = \sum_{i:y_{ij}=1} \log s_{ij}(\theta)$ (but let's not do that).

A Purely Statistical Setup: Parametric Forms

In order to compute the choice probabilities, we must perform a J dimensional integral over $f(\boldsymbol{\varepsilon_i})$.

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

There are some choices that make our life easier

- Multivariate normal: $\varepsilon_i \sim \mathcal{N}(0, \Omega)$. $\longrightarrow$ multinomial probit.
- Gumbel/Type 1 EV: $f(\varepsilon_{ij}) = e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}}$ and $F(\varepsilon_{ij}) = 1 - e^{-e^{-\varepsilon_{ij}}}$ $\longrightarrow$ multinomial logit
- There are also heteroskedastic variants of the Type I EV/ Logit framework.

Parametric Forms: Multinomial Probit

Multinomial Probit?

- The probit has $\varepsilon_i \sim \mathcal{N}(0, \Sigma)$.
- If Σ is unrestricted, then this can produce relatively flexible substitution patterns.
- Flexible is relative: still have normal tails, only pairwise correlations, etc.
- It might be that ρ_{12} is large if $1, 2$ are similar products.
- Much more flexible than Logit

Downside

- Σ has potentially J^2 parameters (that is a lot)!
- Maybe $J * (J-1)/2$ under symmetry. (still a lot).
- Each time we want to compute $s_{ij}(\theta)$ we have to simulate an integral of dimension J.
- I wouldn't do this for $J \geq 5$.

Multinomial Logit

Multinomial Logit has closed form choice probabilities

$$s_{ij} = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

Expected maximum also has closed form:

$$\mathbb{E}[\max_j u_{ij}] = \log\left(\sum_j \exp[V_{ij}]\right) + \gamma$$

Logit Inclusive Value is helpful for several reasons

- Expected utility of best option (without knowledge of ε_i) does not depend on ε_{ij}.
- This is a globally concave function in V_{ij} (more on that later).
- Allows simple computation of ΔCS for consumer welfare (but not CS itself).

Properties of Logit

Multinomial Logit: Identification

What is actually identified here?

- Helpful to look at the ratio of two choice probabilities

$$\log \frac{s_{ij}(\theta)}{s_{ik}(\theta)} = V_{ij} - V_{ik}$$

- We only identify the difference in indirect utilities not the levels.
- This is a feature and not a bug. Why?

Multinomial Logit: Identification

As another idea suppose we add C to each V_{ij}:

$$s_{ij} = \frac{e^{C+V_{ij}}}{\sum_k e^{C+V_{ik}}} = \frac{e^C e^{V_{ij}}}{\sum_k e^C e^{V_{ik}}} = \frac{\exp^{V_{ij}}}{\sum_k \exp^{V_{ik}}}$$

This has no effect. That means we need to fix a normalization C.

- We normalize one of the choices to provide a utility of zero $V_{ik} - \underbrace{V_{ik}}_{=C} = 0$
- We actually already made another normalization. Does anyone know which?

Multinomial Logit: Identification

The most sensible normalization in many settings is to allow for an outside option which produces no utility in expectation.

$$s_{ij} = \frac{e^{V_{ij}}}{1 + \sum_k e^{V_{ik}}}$$

- Hopefully the choice of outside option is well defined: not buying a yogurt, buying some other used car, etc.
- Now this resembles the binomial logit model more closely.

Back to Scale of Utility

- Consider $U^*_{ij} = V_{ij} + \varepsilon^*_{ij}$ with $Var(\varepsilon^*) = \sigma^2\pi^2/6$.
- Without changing behavior we can divide by σ so that $U_{ij} = V_{ij}/\sigma + \varepsilon_{ij}$ and $Var(\varepsilon^*/\sigma) = Var(\varepsilon) = \pi^2/6$

$$s_{ij} = \frac{e^{V_{ij}/\sigma}}{\sum_k e^{V_{ik}/\sigma}} \approx \frac{e^{\beta^*/\sigma \cdot x_{ij}}}{\sum_k e^{\beta^*/\sigma \cdot x_{ik}}}$$

- Every coefficient β is rescaled by σ. This implies that only the ratio β^*/σ is identified.
- Coefficients are relative to variance of unobserved factors. More unobserved variance $\longrightarrow$ smaller β.
- Ratio β_1/β_2 is invariant to the scale parameter σ.

Multinomial Logit: Elasticity

An important output from a demand system are elasticities w.r.t x_{ij}

$$\frac{\partial s_{ij}}{\partial x_{ij}} = (\mathbb{I}[j = k] \cdot s_{ij} - s_{ij} \cdot s_{ik})$$

- This implies that $\eta^x_{jj} = \frac{\partial s_{ij}}{\partial x_{ij}} \frac{x_{ij}}{s_{ij}} = \beta x_{ij}(1 - s_{ij})$.
- $\eta^x_{jk} = \frac{\partial s_{ij}}{\partial x_{ik}} \frac{x_{ik}}{s_{ij}} = -\beta x_{ik} s_{ik}$.
- The own elasticity is increasing in own x_{ij}!
  - Why is this a bad idea? What if x_{ij} is the price of j??
- The cross elasticity doesn't depend on which product j you are talking about!

Multinomial Logit: Substitution

I prefer to characterize substitution via diversion ratios

$$\frac{\partial s_{ij}}{\partial x_{ij}} = (\mathbb{I}[j = k] \cdot s_{ij} - s_{ij} \cdot s_{ik})$$

$$D_{jk} = \frac{\partial s_{ik}}{\partial x_{ij}} \Big/ \left|\frac{\partial s_{ij}}{\partial x_{ij}}\right| = \frac{s_{ik}}{1 - s_{ij}}$$

- Second choices: $\mathbb{P}\left(i \text{ chooses } k \in \mathcal{J} \setminus \{j\} \mid i \text{ chooses } j \in \mathcal{J}\right) = \frac{s_{ik}}{1-s_{ij}}$
- Whether we increase/reduce x_{ij} or remove j from choice set, we get the same substitution patterns.
- Substitution is proportional to share.

Multinomial Logit: Estimation Details

Maximum (log)-likelihood:

$$\max_{\theta} \ell(\theta) = \sum_i \sum_j y_{ij} \log s_{ij}(\theta)$$

Hard part is computing the scores:

$$\frac{\partial \log s_{ij}(\theta)}{\partial \theta} = s_{ij} \cdot (1 - s_{ij}) \cdot \frac{\partial V_{ij}}{\partial \theta}$$

Multinomial Logit: Estimation Details

Method of Moments with some instruments z_{ij}:

$$\sum_i \sum_j \left[y_{ij} - s_{ij}(\theta) \right] z_{ij} = 0$$

It turns out the best choice for z_{ij} is the score (evaluated at true θ_0):

$$z_{ij} = \frac{\partial \log s_{ij}(\theta)}{\partial \theta} = s_{ij} \cdot (1 - s_{ij}) \cdot \frac{\partial V_{ij}}{\partial \theta}$$

Why? this attains the semi-parametric efficiency bound (see Chamberlain 1987).

Multinomial Logit: Estimation Details

Method of Moments with some instruments z_{ij}:

$$\sum_i \sum_j \left[y_{ij} - s_{ij}(\theta) \right] z_{ij} = 0$$

It turns out the best choice for z_{ij} is the score (evaluated at true θ_0):

$$z_{ij} = \frac{\partial \log s_{ij}(\theta)}{\partial \theta} = s_{ij} \cdot (1 - s_{ij}) \cdot \frac{\partial V_{ij}}{\partial \theta}$$

Why? this attains the semi-parametric efficiency bound (see Chamberlain 1987).

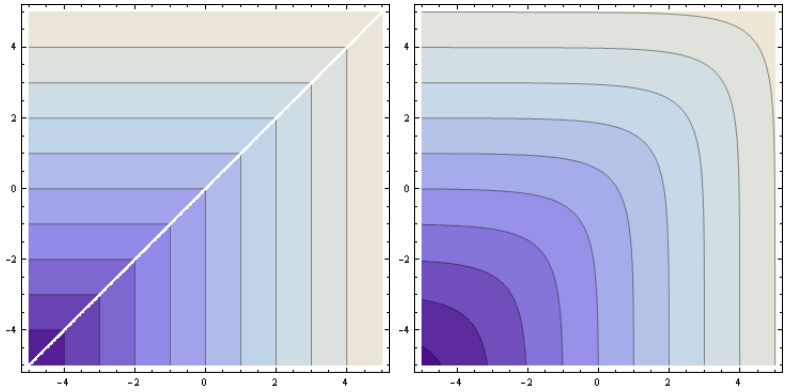
Estimation Details: Softmax

Statistics/Computer Science offer an alternative interpretation of $\log \sum_j \exp$

- Sometimes this is called softmax function .
- Think of this as a continuous/concave approximation to the maximum.
- Consider $\max\{x, y\}$ vs $\log(\exp(x) + \exp(y))$. The exp exaggerates the differences between x and y so that the larger term dominates.
- We can accomplish this by rescaling k: $\log(\exp(kx) + \exp(ky))/k$ as k becomes large the derivatives become infinite and this approximates the "hard" maximum.
- $g(1, 2) = 2.31$, but $g(10, 20) = 20.00004$.

But be careful about how you code this. e^{300} may be too large for your computer!

Estimation Details: Softmax



Thanks!