Multinomial Discrete Choice: IIA Logit

Chris Conlon

April 19, 2020

Applied Econometrics II

Motivation

Most decisions agents make are not necessarily binary:

- Choosing a level of schooling (or a major).
- Choosing an occupation.
- Choosing a partner.
- Choosing where to live.
- Choosing a brand of (yogurt, laundry detergent, orange juice, cars, etc.).

Nonparametric Setup

We consider a multinomial discrete choice:

- in period t
- with J_t alternatives.
- subscript individual agents by i.
- agents choose $j \in J_t$ with probability S_{ijt}.
- Agent i receives utility U_{ij} for choosing j.
- Choice is exhaustive and mutually exclusive.

Nonparametric Setup

We consider a multinomial discrete choice:

- in period t
- with J_t alternatives.
- subscript individual agents by i.
- agents choose $j \in J_t$ with probability S_{ijt}.
- Agent i receives utility U_{ij} for choosing j.
- Choice is exhaustive and mutually exclusive.

Consider the simple example $(t = 1)$:

$$s_{ij} = Prob(U_{ij} > U_{ik} \quad \forall j \neq k)$$

Nonparametric Setup

Now consider separating the utility into the observed V_{ij} and unobserved components ε_{ij}.

$$
\begin{aligned}
s_{ij} &= Prob(U_{ij} > U_{ik} \quad \forall j \neq k) \\
&= Prob(V_{ij} + \varepsilon_{ij} > V_{ik} + \varepsilon_{ik} \quad \forall j \neq k) \\
&= Prob(\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij} \quad \forall j \neq k)
\end{aligned}
$$

Nonparametric Setup

Now consider separating the utility into the observed V_{ij} and unobserved components ε_{ij}.

$$\begin{aligned} s_{ij} &= Prob(U_{ij} > U_{ik} \quad \forall j \neq k) \\ &= Prob(V_{ij} + \varepsilon_{ij} > V_{ik} + \varepsilon_{ik} \quad \forall j \neq k) \\ &= Prob(\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij} \quad \forall j \neq k) \end{aligned}$$

It is helpful to define $f(\varepsilon_i)$ as the J vector of individual i's unobserved utility.

$$\begin{aligned} s_{ij} &= Prob(\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij} \quad \forall j \neq k) \\ &= \int I(\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij}) f(\varepsilon_i) \partial \varepsilon_i \end{aligned}$$

Nonparametric Setup

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

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

There are some choices that make our life easier

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

Errors

Allowing for full support $(-\infty, \infty)$ errors provide two key features:

- Smoothness: s_{ij} is everywhere continuously differentiable in V_{ij}.
- Bound $s_{ij} \in (0, 1)$ so that we can rationalize any observed pattern in the data.
- What does ε_{ij} really mean? (unobserved utility, idiosyncratic tastes, etc.)

Basic Identification

- Only differences in utility matter: $Prob(\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij} \quad \forall j \neq k)$
- Adding constants is irrelevant: if $U_{ij} > U_{ik}$ then $U_{ij} + a > U_{ik} + a$.
- Only differences in alternative specific constants can be identified

$$\begin{aligned} U_b &= X_b\beta + k_b + \varepsilon_b \\ U_c &= X_c\beta + k_c + \varepsilon_c \end{aligned}$$

  only $d = k_b - k_c$ is identified.
- This means that we can only include $J - 1$ such k's and need to normalize one to zero. (Much like fixed effects).
- We cannot have individual specific factors that enter the utility of all options such as income θY_i. We can allow for interactions between individual and choice characteristics $\theta p_j / Y_i$.

Basic Identification: Location

- Technically we can't really fully specify $f(\varepsilon_i)$ since we can always re-normalize: $\widetilde{\varepsilon_{ijk}} = \varepsilon_{ij} - \varepsilon_{ik}$ and write $g(\widetilde{\varepsilon_{ik}})$. Thus any $g(\widetilde{\varepsilon_{ik}})$ is consistent with infinitely many $f(\varepsilon_i)$.
- Logit pins down $f(\varepsilon_i)$ sufficiently with parametric restrictions.
- Probit does not. We must generally normalize one dimension of $f(\varepsilon_i)$ in the probit model. Usually a diagonal term of Ω so that $\omega_{11} = 1$ for example. (Actually we need to do more!).

Basic Identification: Scale

- Consider: $U_{ij}^0 = V_{ij} + \varepsilon_{ij}$ and $U_{ij}^1 = \lambda V_{ij} + \lambda\varepsilon_{ij}$ with $\lambda > 0$. Multiplying by constant λ factor doesn't change any statements about $U_{ij} > U_{ik}$.
- We normalize this by fixing the variance of ε_{ij} since $Var(\lambda\varepsilon_{ij}) = \sigma_e^2\lambda^2$.
- Normalizing this variance normalizes the scale of utility.
- For the logit case the variance is normalized to $\pi^2/6$. (this emerges as a constant of integration to guarantee a proper density).

Observed Heteroskedasticity

Consider the case where $Var(\varepsilon_{ij}^{B}) = \sigma^2$ and $Var(\varepsilon_{ij}^{C}) = k^2\sigma^2$:

- We can estimate

$$\begin{aligned} U_{ij} &= x_j\beta + \varepsilon_{ij}^{B} \\ U_{ij} &= x_j\beta + \varepsilon_{ij}^{C} \end{aligned}$$

  becomes:

$$\begin{aligned} U_{ij} &= x_j\beta + \varepsilon_{ij} \\ U_{ij} &= x_j\beta/k + \varepsilon_{ij} \end{aligned}$$

- Some interpret this as saying that in segment C the unobserved factors are $\hat{k}$ times larger.

Deeper Identification Results

Different ways to look at identification

- Are we interested in non-parametric identification of V_{ij}, specifying $f(\varepsilon_i)$?
- Or are we interested in non-parametric identification of U_{ij}. (Generally hard).
  - Generally we require a large support (special-regressor) or "completeness" condition.
  - Lewbel (2000) does random utility with additively separable but nonparametric error.
  - Berry and Haile (2015) with non-separable error (and endogeneity).

Multinomial Logit

- Multinomial Logit has closed form choice probabilities

$$s_{ij} = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}} \approx \frac{e^{\beta' x_{ij}}}{\sum_k e^{\beta' x_{ik}}}$$

- Approximation arises from the hope that we can approximate $V_{ij} \approx X_{ik}\beta$ with something linear in parameters.

Logit Inclusive Value

Expected maximum also has closed form:

$$E[\max_j U_{ij}] = \log\left(\sum_j \exp[V_{ij}]\right) + C$$

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).

Multinomial Logit

Multinomial Logit goes by a lot of names in various literatures

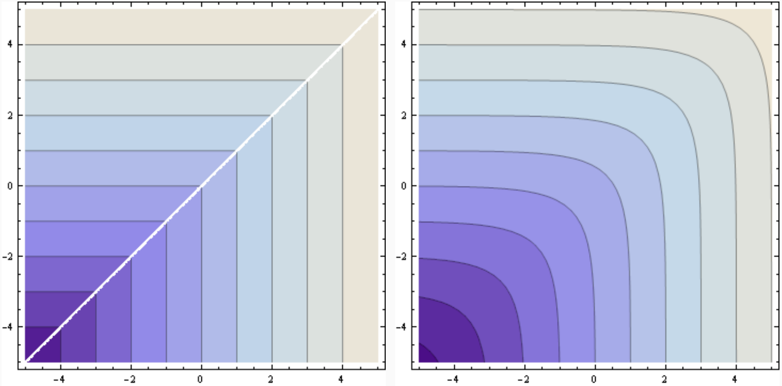
- The problem of multiple choice is often called multiclass classification or softmax regression in other literatures.
- In general these models assume you have individual level data

Alternative Interpretation

Statistics/Computer Science offer an alternative interpretation

- Sometimes this is called softmax regression.
- 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 $g(10, 20) = 20.00004$.

Alternative Interpretation



Multinomial Logit: Identification

What is actually identified here?

- Helpful to look at the ratio of two choice probabilities

$$
\begin{aligned}
\log \frac{s_{ij}(\theta)}{s_{ik}(\theta)} &= \mathbf{x_i}\beta_j - \mathbf{x_i}\beta_k \rightarrow \mathbf{x_i} \cdot (\beta_j - \beta_k) \\
&= \mathbf{x_j}\beta - \mathbf{x_k}\beta \rightarrow (\mathbf{x_j} - \mathbf{x_k}) \cdot \beta
\end{aligned}
$$

- 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 a constant C to each β_j.

$$s_{ij} = \frac{\exp[\mathbf{x_i}(\beta_j + C)]}{\sum_k \exp[\mathbf{x_i}(\beta_k + C)]} = \frac{\exp[\mathbf{x_i}C]\exp[\mathbf{x_i}\beta_j]}{\exp[\mathbf{x_i}C]\sum_k \exp[\mathbf{x_i}\beta_k]}$$

This has no effect. That means we need to fix a normalization C.
The most convenient is generally that $C = -\beta_K$.

- We normalize one of the choices to provide a utility of zero.
- We actually already made another normalization. Does anyone know which?

Multinomial Logit: Identification

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

$$s_{ij} = \frac{\exp[\mathbf{x_i}\beta_j]}{1 + \sum_k \exp[\mathbf{x_i}\beta_k]}$$

- 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.

Taste Variation

- Logit allows for taste variation across individuals if two conditions are met: individual level data and interact observed characteristics only.
- We often want to allow for something like $U_{ij} = x_j\beta_i - \alpha_i p_j + \varepsilon_{ij}$.
- We might want $\beta_i = \theta / y_i$ where y_i is the income for individual i or $\beta_i = \theta y_i$, etc.
- Can also have z_{ij} such as the distance between i and hospital j.
- Cannot have unobserved heterogeneity or heteroskedasticity in ε_{ij}.

Taste Variation

$$\frac{s_{ij}}{s_{ik}} = \frac{e^{V_{ij}}}{\sum_{k'} e^{V_{ik'}}} \Big/ \frac{e^{V_{ik}}}{\sum_{k'} e^{V_{ik'}}} = \frac{e^{V_{ij}}}{e^{V_{ik}}} = \exp[V_{ij} - V_{ik}].$$

- The ratio of choice probabilities for j and k depends only on j and k and not on any alternative l, this is known as independence of irrelevant alternatives.
- For some (Luce (1959)) IIA was an attractive property for axiomatizing choice.
- In fact the logit was derived in the search for a statistical model that satsified various axioms.

IIA Property

- The well known counterexample: You can choose to go to work on a car c or blue bus bb. $S_c = S_{bb} = \frac{1}{2}$ so that $\frac{S_c}{S_{bb}} = 1$.
- Now we introduce a red bus rb that is identical to bb. Then $\frac{S_{rb}}{S_{bb}} = 1$ and $S_c = S_{bb} = S_{rb} = \frac{1}{3}$ as the logit model predicts.
- In reality we don't expect painting a bus red would change the number of individuals who drive a car so we would anticipate $S_c = \frac{1}{2}$ and $S_{bb} = S_{rb} = \frac{1}{4}$.
- We may not encounter too many cases where $\rho_{\varepsilon_{ik},\varepsilon_{ij}} \approx 1$, but we have many cases where this $\rho_{\varepsilon_{ik},\varepsilon_{ij}} \neq 0$
- What we need is the ratio of probabilities to change when we introduce a third option!

IIA Property

- IIA implies that we can obtain consistent estimates for β on any subset of alternatives.
- This means instead of using all $\mathcal{J}$ alternatives in the choice set, we could estimate on some subset $\mathcal{S} \subset \mathcal{J}$.
- This used to be a way to reduce the computational burden of estimation (not clear this is an issue in 2016).
- Sometimes we have choice based samples where we oversample people who choose a particular alternative. Manski and Lerman (1977) show we can get consistent estimates for all but the ASC. This requires knowledge of the difference between the true rate A_j and the choice-based sample rate $\mathcal{S}_j$.
- Hausman proposes a specification test of the logit model: estimate on the full dataset to get $\hat{\beta}$, construct a smaller subsample $\mathcal{S}^k \subset \mathcal{J}$ and $\hat{\beta}^k$ for one or more subsets k. If $|\hat{\beta}^k - \hat{\beta}|$ is small enough.

IIA Property

For the linear V_{ij} case we have that $\frac{\partial V_{ij}}{\partial z_{ij}} = \beta_z$.

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

And Elasticity: $$\frac{\partial \log s_{ij}}{\partial \log z_{ij}} = s_{ij}(1 - s_{ij})\frac{\partial V_{ij}}{\partial z_{ij}}\frac{z_{ij}}{s_{ij}} = (1 - s_{ij})z_{ij}\frac{\partial V_{ij}}{\partial z_{ij}}$$

With cross effects: $$\frac{\partial s_{ij}}{\partial z_{ik}} = -s_{ij}s_{ik}\frac{\partial V_{ik}}{\partial z_{ik}}$$

and elasticity : $$\frac{\partial \log s_{ij}}{\partial \log z_{ik}} = -s_{ik}z_{ik}\frac{\partial V_{ik}}{\partial z_{ik}}$$

Proportional Substitution

Cross elasticity doesn't really depend on j.

$$\frac{\partial \log s_{ij}}{\partial \log z_{ik}} = -s_{ik} z_{ik} \underbrace{\frac{\partial V_{ik}}{\partial z_{ik}}}_{\beta_z}.$$

- This leads to the idea of proportional substitution. As option k gets better it proportionally reduces the shares of the all other choices.
- Likewise removing an option k means that $\tilde{s}_{ij} = \frac{s_{ij}}{1-s_{ik}}$ for all other j.
- This might be a desirable property but probably not.

Multinomial Logit: Estimation with Individual Data

Estimation is straightforward via Maximum Likelihood (MLE):

$$
\begin{aligned}
L(\mathbf{y}|\mathbf{x},\theta) &= \prod_{i=1}^{N} \underbrace{\frac{n_i!}{\prod_{j=1}^{J} y_{ij}!}}_{C(\mathbf{y})} \prod_{j=1}^{J} s_{ij}(x_{ij},\theta)^{y_{ij}} \\
ll(\mathbf{y}|\mathbf{x},\theta) &= \sum_{i=1}^{N} \log(C(\mathbf{y})) + \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\log(s_{ij}(x_{ij},\theta)) \\
l(\mathbf{y}|\mathbf{x},\theta) &\approx \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\log(s_{ij}(x_{ij},\theta))
\end{aligned}
$$

- We can ignore the combinatorial term (with the factorials) since it does not affect the location of the maximum (it is additive and doesn't depend on θ).

Multinomial Logit: Inclusive Value

To be more specific:

- Let's look a little more closely at what's going on:

$$\sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\left[x_{ij}\beta - \underbrace{\log\left(\sum_{k=1}^{K} x_{ik}\beta\right)}_{IV_i(\mathbf{x_i},\theta)}\right]$$

- We call the term on the right the logit inclusive value. It does not depend on k but might vary across choice situations/individuals i.
- The point of the inclusive value is to guarantee that $\sum_{k=1} s_{ik}(\mathbf{x_i},\theta) = 1$.
- If we somehow observed $IV_i(\theta)$ we could just do linear regression (in fact we could do this separately for each K).

Multinomial Logit: Estimation with Aggregate Data

Estimation is just like before

- Suppose that all consumers had the same $x_{ij} = x_j$ (Choices depended only on products not on income, education, etc.)
- We can construct $y_j^* = \sum_{i=1}^N y_{ij}$.

$$l(\mathbf{y}|\mathbf{x},\theta) \approx \sum_{j=1}^{J} y_j^* \log(s_j(\mathbf{x},\theta))$$

- When each consumer i faces the same choice environment, we can aggregate data into sufficient statistics.

Multinomial Logit: Estimation with Aggregate Data

Aggregation is probably the most important property of discrete choice:

- Instead of individual data, or a single group we might have multiple groups: if prices only change once per week, we can aggregate all of the week's sales into one "observation".
- Likewise if we only observe that an individual is within one of five income buckets – there is no loss from aggregating our data into these five buckets.
- All of this depends on the precise form of $s_j(\mathbf{x_i}, \theta)$. When it doesn't change across observations: we can aggregate.
- It functions as if we have a representative consumer up to ε_i.
- We can use this idea to go from individual level to market demand: $q_j(\mathbf{x_i}) = N_i s_{ij}(\theta)$.

Multinomial Logit: Elasticity

An important output from a demand system are elasticities

- An important element in $\mathbf{x_i}$ are prices $[p_1, \ldots, p_J]$
- Helpful to write $u_{ij} = x_j\beta - \alpha p_j$ (assumes aggregation!).

$$\frac{\partial q_j}{\partial p_k} = -N \cdot \alpha \left(I[j = k]s_j - s_j s_k\right)$$

- This implies that $\eta_{jj} = \frac{\partial q_j}{\partial p_j}\frac{p_j}{q_j} = -\alpha p_j(1 - s_j)$.
- The price elasticity is increasing in own price! (Why is this a bad idea?)
- $\eta_{jk} = \frac{\partial q_j}{\partial p_k}\frac{p_k}{q_j} = -\alpha p_k s_k$.
- The cross price elasticity doesn't depend on which product j you are talking about!

Thanks!