Multinomial Discrete Choice: IIA Logit

Chris Conlon

Fall 2025

Grad IO

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 $\mathcal{J}_t$ alternatives.
- subscript individual agents by i.
- agents choose $j \in J_t$ with probability s_{ijt}.
- Agent i receives utility U_{ijt} for choosing j.
- Choice is exhaustive and mutually exclusive.

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

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

Nonparametric Setup

We consider a multinomial discrete choice:

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

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

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

Nonparametric Setup

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

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

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

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

Nonparametric Setup

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

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

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

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

Nonparametric Setup

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

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

There are some choices that make our life easier

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

Errors

Allowing for a continuous density with 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.
  - Caveat: zero and one (interpretation).
- What does ε_{ij} really mean? (unobserved utility, idiosyncratic tastes, etc.)

Basic Identification

- Only differences in utility matter: $\mathbb{P}(\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij} \quad \forall k \neq j)$
- 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

$$U_b = v_{ib} + k_b + \varepsilon_{ib}$$
$$U_c = v_{ic} + k_c + \varepsilon_{ic}$$

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 $\theta p_j / Y_i$.

$$U_b = v_b + \theta y_i + \varepsilon_b$$
$$U_c = v_c + \theta y_i + \varepsilon_c$$

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 $Var(\varepsilon_{ib}) = \sigma^2$ and $Var(\varepsilon_{ic}) = k^2\sigma^2$:

- We can estimate

$$\begin{aligned} U_{ib} &= v_{ib} + \varepsilon_{ib} \\ U_{ic} &= v_{ic} + \varepsilon_{ic} \end{aligned}$$

  becomes:

$$\begin{aligned} U_{ib} &= v_{ib} + \varepsilon_{ib} \\ U_{ic} &= v_{ic} + \varepsilon_{ic} \end{aligned}$$

- 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 (Gumbel/Type I EV) has closed form choice probabilities

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

- Often we approximate $V_{ij} \approx X_{ik}\beta$ with something linear in parameters.

Logit Inclusive Value

Expected maximum also has closed form:

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

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

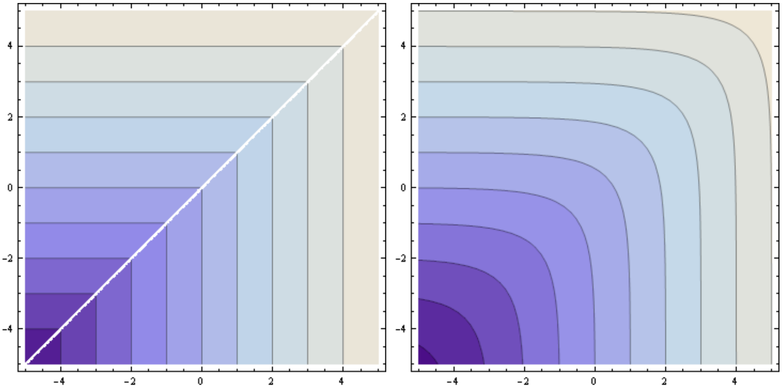
Multinomial Logit goes by a lot of names in various literatures

- 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 $g(10, 20) = 20.00004$.


4
2
0
-2
-4
-4
-2
0
2
4
4
2
0
-2
-4
-4
-2
0
2
4

Multinomial Logit: Identification

What is actually identified here?

- Helpful to look at the ratio of two choice probabilities

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

- We only identify the difference in indirect utilities not the levels.
- 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. (A feature or a bug?)
- In fact the logit was derived in the search for a statistical model that satsified various axioms.

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 so that $e^{V_{i0}} = e^0 = 1$:

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

- 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. (marginal rate of substitution).

IIA Property

The well known critique:

- 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 21st century).
- 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 : $$\frac{\partial \log s_{ij}}{\partial \log z_{ik}} = -s_{ik}z_{ik}\frac{\partial V_{ik}}{\partial z_{ik}}$$

Own and Cross Elasticity

An important output from a demand system are elasticities

- This implies that $\eta_{jj} = \frac{\partial s_{ij}}{\partial p_j} \frac{p_j}{s_{ij}} = \beta_p \cdot p_j \cdot (1 - s_{ij})$.
- The price elasticity is increasing in own price! (Why is this a bad idea?)
- Also mechanical relationship between elasticity and share so that popular products necessarily have higher markups (holding fixed prices).

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.
- This might be a desirable property but probably not.

Diversion Ratios

Recall the diversion ratio:

$$D_{jk} = \frac{\frac{\partial s_{ik}}{\partial p_j}}{\left|\frac{\partial s_{ij}}{\partial p_j}\right|} = \frac{\beta_p s_{ik} s_{ij}}{\beta_p s_{ij}(1 - s_{ij})} = \frac{s_{ik}}{1 - s_{ij}}$$

- Again proportional substitution. As price of j goes up we proportionally inflate choice probabilities of substitutes.
- Likewise removing an option j means that $\tilde{s}_{ik}(\mathcal{J} \backslash j) = \frac{s_{ik}}{1 - s_{ij}}$ for all other k.
- IIA/Logit means constant diversion ratios.

Thanks!