

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 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 ($t = 1$):

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

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_{ijt} for choosing j .

Choice is exhaustive and mutually exclusive.

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

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

Nonparametric Setup

Now consider separating the utility into the **observed** V_{ij} and **unobserved** components $_{ij}$.

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

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

$$\begin{aligned}s_{ij} &= \Pr(_{ij} - _{ik} > V_{ik} - V_{ij} \quad k \neq j) \\ &= \mathbb{I}(_{ij} - _{ik} > V_{ik} - V_{ij})f(i)_i\end{aligned}$$

Nonparametric Setup

Now consider separating the utility into the **observed** V_{ij} and **unobserved** components $_{ij}$.

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

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

$$\begin{aligned}s_{ij} &= \Pr(_{ij} - _{ik} > V_{ik} - V_{ij} \quad k \neq j) \\ &= I(_{ij} - _{ik} > V_{ik} - V_{ij})f(i)_i\end{aligned}$$

Nonparametric Setup

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

$$s_{ij} = \int I(v_{ij} - v_{ik} > 0) f(i) di$$

There are some choices that make our life easier

Multivariate normal: $i \sim N(0, \Sigma)$. **multinomial probit.**

Gumbel/Type 1 EV: $f(i) = \frac{1}{J} \prod_{j=1}^J e^{-v_{ij}}$ and $F(i) = \frac{1}{J} \sum_{j=1}^J e^{-v_{ij}}$ **multinomial logit**

There are also heteroskedastic variants of the Type I EV/ Logit framework.

Allowing for a continuous density with full support $(,)$ 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 s_{ij} really mean? (unobserved utility, idiosyncratic tastes, etc.)

Basic Identification

Only differences in utility matter: $\Pr(ij > ik) = \Pr(V_{ij} - V_{ik} > 0)$

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

$$U_b = v_{ib} + k_b + \epsilon_{ib}$$

$$U_c = v_{ic} + k_c + \epsilon_{ic}$$

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 Y_i . We can allow for interactions between individual and choice characteristics p_j/Y_i .

$$U_b = v_b + y_i + \epsilon_{ib}$$

$$U_c = v_c + y_i + \epsilon_{ic}$$

Basic Identification: Location

Technically we can't really fully specify $f(i)$ since we can always re-normalize: $\pi_{ijk} = \pi_{ij} \pi_{ik}$ and write $g(\pi_{ik})$. Thus any $g(\pi_{ik})$ is consistent with infinitely many $f(i)$.

Logit pins down $f(i)$ sufficiently with parametric restrictions.

Probit does not. We must generally normalize one dimension of $f(i)$ in the probit model. Usually a diagonal term of so that $\pi_{11} = 1$ for example. (Actually we need to do more!).

Basic Identification: Scale

Consider: $U_{ij}^0 = V_{ij} + \epsilon_{ij}$ and $U_{ij}^1 = V_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} > 0$. Multiplying by constant factor doesn't change any statements about $U_{ij} > U_{ik}$.

We normalize this by fixing the variance of ϵ_{ij} since $Var(\epsilon_{ij}) = \frac{2^2}{e}$.

Normalizing this variance normalizes the scale of utility.

For the logit case the variance is normalized to $2^2/6$. (this emerges as a constant of integration to guarantee a proper density).

Observed Heteroskedasticity

Consider the case where $Var(u_{ib}) = \sigma^2$ and $Var(u_{ic}) = k^2\sigma^2$:

We can estimate

$$U_{ib} = v_{ib} + u_{ib}$$

$$U_{ic} = v_{ic} + u_{ic}$$

becomes:

$$U_{ib} = v_{ib} + u_{ib}$$

$$U_{ic} = v_{ic} + u_{ic}$$

Some interpret this as saying that in segment C the unobserved factors are k times larger.

Deeper Identification Results

Different ways to look at identification

Are we interested in non-parametric identification of V_{ij} , specifying $f(i)$?

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 $V_{ij} = X_{ijk} \beta_k$ 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 i) does not depend on i_j .

- This is a globally concave function in V_{ij} (more on that later).

- Allows simple computation of CS 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.

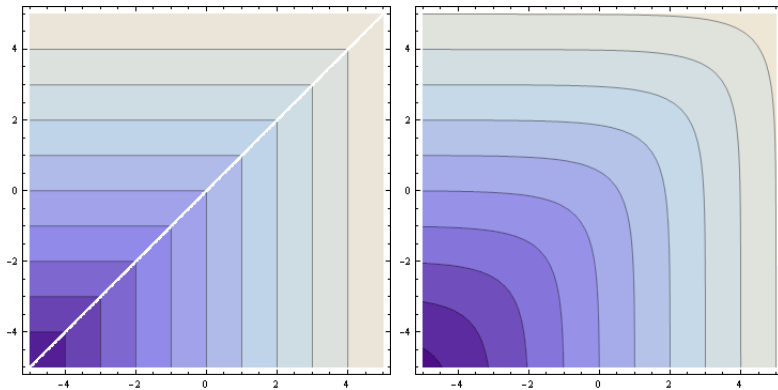
Think of this as a continuous/concave approximation to the maximum.

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

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

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 satisfied various axioms.

Multinomial Logit: Identification

As another idea suppose we add a constant C to each j .

$$s_{ij} = \frac{\exp[x_i(j + C)]}{\sum_k \exp[x_i(k + C)]} = \frac{\exp[x_i C] \exp[x_i j]}{\exp[x_i C] \sum_k \exp[x_i k]}$$

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

The most convenient is generally that $C = -\ln \sum_k \exp[x_i k]$.

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 $U_{ij} = V_{ij} + \epsilon_{ij}$ with $Var(\epsilon_{ij}) = \sigma^2/6$.

Without changing behavior we can divide by σ so that $U_{ij} = V_{ij}/\sigma + \epsilon_{ij}$ and $Var(\epsilon_{ij}/\sigma) = Var(\epsilon_{ij})/\sigma^2 = 1/6$

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

Every coefficient β_j is rescaled by σ . This implies that only the ratio β_j/β_1 is identified.

Coefficients are relative to variance of unobserved factors. More unobserved variance smaller σ .

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 $\frac{S_{ik}}{S_{ij}} = 1$, but we have many cases where this

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 θ_j on any subset of alternatives.

This means instead of using all \mathcal{J} alternatives in the choice set, we could estimate on some subset $S \subseteq \mathcal{J}$.

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 S_j .

Hausman proposes a specification test of the logit model: estimate on the full dataset to get θ_j , construct a smaller subsample $S^k \subseteq \mathcal{J}$ and θ_j^k for one or more subsets k . If $|\theta_j - \theta_j^k|$ is small enough.

IIA Property

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

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

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

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

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

Own and Cross Elasticity

An important output from a demand system are elasticities

This implies that $\epsilon_{jj} = \frac{s_{ij}}{p_j} \frac{p_j}{s_{ij}} = \epsilon_p (1 - s_{ij})$.

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 .

$$\frac{\log s_{ij}}{\log z_{ik}} = s_{ik} z_{ik} \frac{V_{ik}}{z_{ik}} .$$

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:

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

Again proportional substitution. As price of j goes up we proportionally inflate choice probabilities of substitutes.

Likewise removing an option j means that $s_{ik}(\mathcal{J} - j) = \frac{s_{ik}}{1 - s_{ij}}$ for all other k .

IIA/Logit means **constant diversion ratios**.

Thanks!
