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 $j \quad J_t$ with probability s_{ijt}.

Agent i receives utility U_{ijt} for choosing j.

Choice is exhaustive and mutually exclusive.

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

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

Nonparametric Setup

We consider a multinomial discrete choice:

in period t

with $\mathcal{J}_t$ alternatives.

subscript individual agents by i.

agents choose $j \quad J_t$ with probability s_{ijt}.

Agent i receives utility U_{ijt} for choosing j.

Choice is exhaustive and mutually exclusive.

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

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

Nonparametric Setup

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

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

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

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

Nonparametric Setup

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

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

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

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

Nonparametric Setup

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

$$s_{ij} = I({}_{ij}\ \ {}_{ik} > V_{ik}\ \ V_{ij}) f({}_i) {}_i$$

There are some choices that make our life easier

Multivariate normal: ${}_i\ \ N(0,)$. multinomial probit.

Gumbel/Type 1 EV: $f({}_i) = e^{{}_{ij}} e^{e^{{}_{ij}}}$ and $F({}_i) = 1\ \ e^{e^{{}_{ij}}}$ multinomial logit

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

Errors

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} $(0, 1)$ so that we can rationalize any observed pattern in the data.
  - Caveat: zero and one (interpretation).
- What does $_{ij}$ really mean? (unobserved utility, idiosyncratic tastes, etc.)

Basic Identification

Only differences in utility matter: $\Pr({}_{ij}\ {}_{ik} > V_{ik}\ V_{ij}\quad k\ j)$

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

$$U_b = v_{ib} + k_b + {}_{ib}$$

$$U_c = v_{ic} + k_c + {}_{ic}$$

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.

$$U_b = v_b + y_i + {}_b$$

$$U_c = v_c + y_i + {}_c$$

Basic Identification: Location

Technically we can't really fully specify $f(_i)$ since we can always re-normalize: $_{ijk} = _{ij} \quad _{ik}$ and write $g(_{ik})$. Thus any $g(_{ik})$ 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 $_{11} = 1$ for example. (Actually we need to do more!).

Basic Identification: Scale

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

We normalize this by fixing the variance of ${}_{ij}$ since $Var({}_{ij}) = {}^{22}_{e}$.

Normalizing this variance normalizes the scale of utility.

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

Observed Heteroskedasticity

Consider the case where $Var({}_{ib}) = {}^2$ and $Var({}_{ic}) = k^{22}$:

We can estimate

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

becomes:

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

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

$$s_{ij} = \frac{e^{V_{ij}}}{{}_{k}\, e^{V_{ik}}}$$

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

Logit Inclusive Value

Expected maximum also has closed form:

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

Logit Inclusive Value is helpful for several reasons

Expected utility of best option (without knowledge of $_i$) does not depend on $_{ij}$.

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.

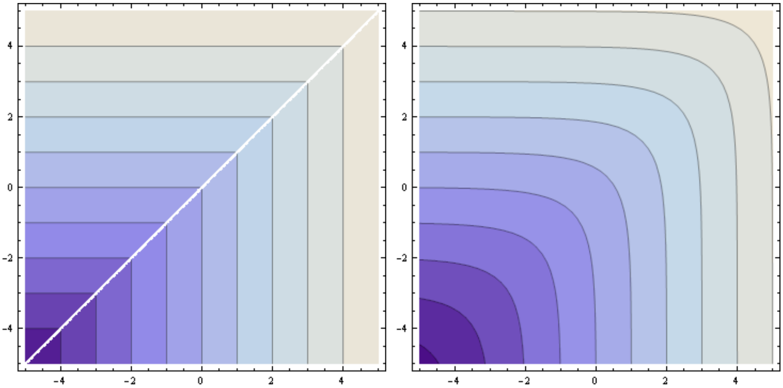
Think of this as a continuous/concave approximation to the maximum.

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

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

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 $_j$.

$$s_{ij} = \frac{\exp[\mathrm{x_i}(_j + C)]}{_k \exp[\mathrm{x_i}(_k + C)]} = \frac{\exp[\mathrm{x_i}C]\exp[\mathrm{x_i}{}_j]}{\exp[\mathrm{x_i}C]\,{}_k \exp[\mathrm{x_i}{}_k]}$$

This has no effect. That means we need to fix a normalization C.
The most convenient is generally that $C = {}_K$.

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

$$s_{ij} = \frac{e^{V_{ij}}}{1 + {}_{k}\, e^{V_{ik}}}$$

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 $U_{ij} = V_{ij} + {}_{ij}$ with $Var() = {}^{22}/6$.

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

$$s_{ij} = \frac{e^{V_{ij}/}}{{}_{k}\, e^{V_{ik}/}} \quad \frac{e^{/x_{ij}}}{{}_{k}\, e^{/x_{ik}}}$$

Every coefficient is rescaled by . This implies that only the ratio $/$ is identified.

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

Ratio ${}_1/{}_2$ is invariant to the scale parameter . (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 ${}_{ik,ij}\ \ 1$, but we have many cases where this ${}_{ik,ij}\ \ 0$

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 on any subset of alternatives.

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

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 , construct a smaller subsample $\mathcal{S}^k$ $\mathcal{J}$ and ^{k} for one or more subsets k. If $|^k \quad |$ is small enough.

IIA Property

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

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

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

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 $_{jj} = \frac{s_{ij}}{p_j}\frac{p_j}{s_{ij}} = {}_p \; p_j \; (1 \;\; s_{ij})$.

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.

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

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:

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

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!