Binary/Multinomial Choice and Statistical Demand Models*

Charlie Murry

Empirical Methods | University of Michigan

February 5, 2025

**Based heavily on notes from Chris Conlon (NYU Stern), Rich Sweeney (BC) and others.*

Multinomial Choice

Motivation

Most decisions agents make are not necessarily binary:

- Choosing a level of schooling (or a major).
- Choosing an occupation.
- Choosing a partner.
- Choosing a mutual fund/manager.
- Choosing where to live.
- Choosing a brand/model 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 P_{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 P_{ijt}.
- Agent i receives utility U_{ij} for choosing j.
- Choice is exhaustive and mutually exclusive.

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

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

Nonparametric Setup

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

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

Notice additive separability.

Nonparametric Setup

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

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

Notice additive separability.

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

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

Nonparametric Setup

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

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

There are some choices that make our life easier

- Multivariate normal: $\varepsilon_i \sim N(0, \Omega)$. $\longrightarrow$ 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

Errors

Allowing for full support $[-\infty, \infty]$ errors provide two key features:

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

Basic Identification

- Only differences in utility matter: $\mathrm{Prob}(\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij} \quad \forall j \neq k)$
- 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 characterstics $\theta p_j / Y_i$.

Basic Identification

Location

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

- e.g. we only observe if $\mathrm{U}_{\mathrm{ij}} > \mathrm{U}_{\mathrm{ik}}$, not by how much.

- Logit pins down $\mathrm{f}(\varepsilon_{\mathrm{i}})$ sufficiently with parametric restrictions (no covariance terms).

- Probit does not. We must generally normalize one dimension of $\mathrm{f}(\varepsilon_{\mathrm{i}})$ in the probit model. Usually a diagonal term of Ω so that $\omega_{11} = 1$ for example.

Basic Identification

Scale

- Consider: $\mathrm{U}_{\mathrm{ij}}^{0} = \mathrm{V}_{\mathrm{ij}} + \varepsilon_{\mathrm{ij}}$ and $\mathrm{U}_{\mathrm{ij}}^{1} = \lambda \mathrm{V}_{\mathrm{ij}} + \lambda \varepsilon_{\mathrm{ij}}$ with $\lambda > 0$. Multiplying by constant λ factor doesn't change any statements about $\mathrm{U}_{\mathrm{ij}} > \mathrm{U}_{\mathrm{ik}}$.

- We normalize this by fixing the variance of $\varepsilon_{\mathrm{ij}}$ since $\mathrm{Var}(\lambda \varepsilon_{\mathrm{ij}}) = \sigma_{\mathrm{e}}^{2} \lambda^{2}$.

- 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).

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 (2024, ECMA) with non-separable error (and endogeneity).

Multinomial Logit (MNL)

- 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.
- Expected maximum also has closed form:

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

Logit Inclusive Value

- Logit Inclusive Value is helpful for several reasons

$$\mathrm{E}[\max_{\mathrm{j}} \mathrm{U}_{\mathrm{ij}}] = \log\left(\sum_{\mathrm{j}} \exp[\mathrm{V}_{\mathrm{ij}}]\right) + \mathrm{C}$$

- Expected utility of best option (no knowledge of realized 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 $\Delta\mathrm{CS}$ for consumer welfare.

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

Multinomial Logit: Identification

What is actually identified here?

- Helpful to look at the ratio of two choice probabilities

$$\log \frac{\mathrm{s_{ij}}(\theta)}{\mathrm{s_{ik}}(\theta)} = \mathrm{x_{ij}}\beta_{\mathrm{j}} - \mathrm{x_{ik}}\beta_{\mathrm{k}} \rightarrow \mathrm{x_i} \cdot (\beta_{\mathrm{j}} - \beta_{\mathrm{k}})$$

- We only identify the difference in indirect utilities not the levels.

Multinomial Logit: Identification

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

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

- This has no effect. That means we need to fix a normalization C. The most convenient is generally that $\mathrm{C} = -\beta_{\mathrm{K}}$.
- We normalize one of the choices to provide a utility of zero.
- This is on top of the scale normalization. But these don't impact the behavioral implications of the model.

Multinomial Logit: Identification

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

$$\mathrm{s_{ij}} = \frac{\exp[\mathrm{x_i}\beta_\mathrm{j}]}{1 + \sum_\mathrm{k} \exp[\mathrm{x_i}\beta_\mathrm{k}]}$$

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

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

- 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.
- In fact, if price is a regressor, we can scale other parameters by price coefficient to get utils in terms of $.

Taste Variation

- Logit allows for taste variation across individuals if two conditions are met: individual level data and interact observed characterstics 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

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

- 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. $\mathrm{P_c} = \mathrm{P_{bb}} = \frac{1}{2}$ so that $\frac{\mathrm{P_c}}{\mathrm{P_{bb}}} = 1$.
- Now we introduce a red bus rb that is identical to bb. Then $\frac{\mathrm{P_{rb}}}{\mathrm{P_{bb}}} = 1$ and $\mathrm{P_c} = \mathrm{P_{bb}} = \mathrm{P_{rb}} = \frac{1}{3}$ 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 $\mathrm{P_c} = \frac{1}{2}$ and $\mathrm{P_{bb}} = \mathrm{P_{rb}} = \frac{1}{4}$.
- We may not encounter too many cases where $\rho_{\varepsilon_{\mathrm{ik}},\varepsilon_{\mathrm{ij}}} \approx 1$, but we have many cases where this $\rho_{\varepsilon_{\mathrm{ik}},\varepsilon_{\mathrm{ij}}} \neq 0$
- 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 J alternatives in the choice set, we could estimate on some subset $S \subset J$.
- 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 constant. 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 $\hat{\beta}$, construct a smaller subsample $S^k \subset J$ and $\hat{\beta}^k$ for one or more subsets k. If $|\hat{\beta}^k - \hat{\beta}|$ is small enough.

IIA Property

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

And Elasticity:

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

With cross effects:

$$\frac{\partial \mathrm{s_{ij}}}{\partial \mathrm{z_{ik}}} = -\mathrm{s_{ij}}\mathrm{s_{ik}}\frac{\partial \mathrm{V_{ik}}}{\partial \mathrm{z_{ik}}}$$

And Elasticity:

$$\frac{\partial \log \mathrm{s_{ij}}}{\partial \log \mathrm{z_{ik}}} = -\mathrm{s_{ik}}\mathrm{z_{ik}}\frac{\partial \mathrm{V_{ik}}}{\partial \mathrm{z_{ik}}}$$

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

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 is not a desireable property for most empirical work.

Multinomial Logit: Estimation with Individual Data

Estimation is straightforward via Maximum Likelihood (MLE):

$$\mathrm{l}(\mathrm{y}|\mathrm{x},\theta) \approx \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{y}_{\mathrm{ij}} \log(\mathrm{s}_{\mathrm{ij}}(\mathrm{x}_{\mathrm{ij}},\theta))$$

This ends up being many terms if there are many people and many choices in the dataset. But it's not really a problem for modern computers. Plus, the problem is convex.

Multinomial Logit: Inclusive Value

To be more specific:

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

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

- 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 $\sum_{k=1} s_{ik}(x_i,\theta) = 1$.
- 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 $y_j^* = \sum_{i=1}^{N} y_{ij}$.

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

- 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 the logit:

- 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 $s_j(x_i, \theta)$. 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: $q_j(x_i) = N_i s_{ij}(\theta)$.

Multinomial Logit: Elasticitiy

An important output from a demand system are elasticities

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

$$\frac{\partial q_j}{\partial p_k} = -N \cdot \alpha \left(I[j = k] s_j - \sum_{k=1}^{K} s_j s_k \right)$$

- 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! (no income effects...)
- $\eta_{jk} = \frac{\partial q_j}{\partial p_k} \frac{p_k}{q_j} = -\alpha p_k s_k$.
- The cross price elasiticity doesn't depend on which product j you are talking about!

Multinomial Logit: IIA

The multinomial logit is frequently criticized for producing unrealistic substitution patterns

- Suppose we got rid of a product k then $s_j^{(1)} = s_j^{(0)} \frac{1}{1-s_k}$.
- Substitution is just proportional to your pre-existing shares s_j
- No concept of “closeness” of competition!

Can we do better?

Multinomial Probit?

- The probit has $\varepsilon_i \sim N(0, \Sigma)$.
- 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 $s_j(\theta)$ we have to simulate an integral of dimension J.
- I wouldn't do this for $J \geq 5$.

Relaxing IIA

Let's make ε_{ij} more flexible than IID. Hopefully still have our integrals work out.

$$u_{ij} = x_{ij}\beta + \varepsilon_{ij}$$

- One approach is to allow for a block structure on ε_{ij} (and consequently on the elasticities).
- We assign products into groups g and add a group specific error term

$$u_{ij} = x_{ij}\beta + \eta_g + \varepsilon_{ij}$$

- The trick putting a distribution on $\eta_g + \varepsilon_{ij}$ so that the integrals still work out.
- Do not try this at home: it turns out the required distribution is known as GEV and the resulting model is known as the nested logit.

Nested Logit

A traditional (and simple) relaxation of the IIA property is the Nested Logit. This model is often presented as two sequential decisions.

- First consumers choose a category (following an IIA logit).
- Within a category consumers make a second decision (following the IIA logit).
- This leads to a situation where while choices within the same nest follow the IIA property (do not depend on attributes of other alternatives) choices among different nests do not!

Alternative Interpretation

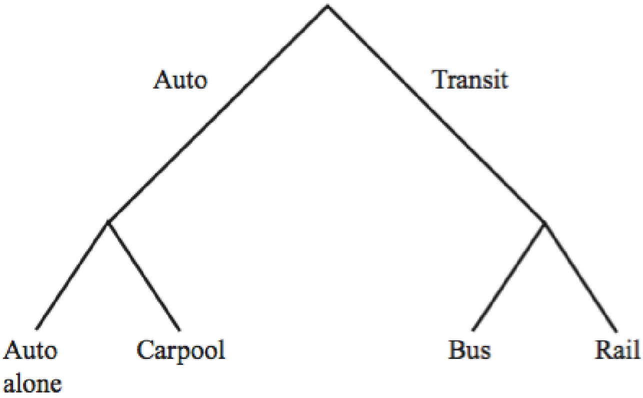

Figure 4.1. Tree diagram for mode choice.

Nested Logit

Utility looks basically the same as before:

$$U_{ij} = V_{ij} + \underbrace{\eta_{ig} + \widetilde{\varepsilon_{ij}}}_{\varepsilon_{ij}(\lambda_g)}$$

- We add a new term that depends on the group g but not the product j and think about it as varying unobservably over individuals i just like ε_{ij}.
- Now $\varepsilon_i \sim F(\varepsilon)$ where $F(\varepsilon) = \exp[-\sum_{g=G}^{G} \left(\sum_{j \in J_g} \exp[-\varepsilon_{ij}/\lambda_g]\right)^{\lambda_g}$. This is no longer Type I EV but GEV.
- The key is the addition of the λ_g parameters which govern (roughly) the within group correlation.
- This distribution is a bit cooked up to get a closed form result, but for $\lambda_g \in [0,1]$ for all g it is consistent with random utility maximization.

Nested Logit

The nested logit choice probabilities are:

$$P_{ij} = \frac{e^{V_{ij}/\lambda_g} \left(\sum_{k \in J_g} e^{V_{ik}/\lambda_g} \right)^{\lambda_g - 1}}{\sum_{h=1}^{G} \left(\sum_{k \in J_h} e^{V_{ik}/\lambda_h} \right)^{\lambda_h}}$$

Within the same group g we have IIA and proportional substitution

$$\frac{P_{ij}}{P_{ik}} = \frac{e^{V_{ij}/\lambda_g}}{e^{V_{ik}/\lambda_g}}$$

But for different groups we do not:

$$P_{ij} = \frac{e^{V_{ij}/\lambda_g} \left(\sum_{k \in J_g} e^{V_{ik}/\lambda_g} \right)^{\lambda_g - 1}}{e^{V_{ik}/\lambda_h} \left(\sum_{k \in J_h} e^{V_{ik}/\lambda_h} \right)^{\lambda_h - 1}}$$

Nested Logit

We can take the probabilities and re-write them slightly with the substitution that $\underbrace{\lambda_g \cdot \log\left(\sum_{k \in J_g} e^{V_{ik}}\right)}_{IV_{ig}}$.

$$
\begin{aligned}
P_{ij} &= \frac{e^{V_{ij}/\lambda_g}}{\left(\sum_{k \in J_g} e^{V_{ik}/\lambda_g}\right)} \cdot \frac{\left(\sum_{k \in J_g} e^{V_{ik}/\lambda_g}\right)^{\lambda_g}}{\sum_{h=1}^{G}\left(\sum_{k \in J_h} e^{V_{ik}/\lambda_h}\right)^{\lambda_h}} \\
&= \underbrace{\frac{e^{V_{ij}/\lambda_g}}{\left(\sum_{k \in J_g} e^{V_{ik}/\lambda_g}\right)}}_{P_{ij|g}} \cdot \underbrace{\frac{e^{\lambda_g IV_{ig}}}{\sum_{h=1}^{G} e^{\lambda_h IV_{ih}}}}_{P_{ig}}
\end{aligned}
$$

This is the decomposition into two logits that leads to the “sequential logit” story.

Nested Logit : Notes

- $\lambda_g = 1$ is the simple logit case (IIA)
- $\lambda_g \to 0$ implies that all consumers stay within the nest.
- $\lambda < 0$ or $\lambda > 1$ can happen and usually means something is wrong. These models are not generally consistent with RUM. (If you report one in your paper I will reject it).
- λ is often interpreted as a correlation parameter and this is almost true but not exactly!
- There are other extensions: overlapping nests, or three level nested logit.
- In general the hard part is understanding what the appropriate nesting structure is ex ante. Maybe for some problems this is obvious but for many not.
- The othr hard part is figuring out what identifies λ

Nested Logit

In practice we end up with the following:

$$s_{ij} = s_{ij|g}(\theta) s_{ig}(\theta)$$

- Because the nested logit can be written as the within group share $s_{ij|g}$ and the share of the group s_{ig} we often explain this model as sequential choice
- First you pick a category, then you pick a product within a category.
- This is a sometimes helpful (sometimes unhelpful) way to think about this.
- We can also think about this as putting a block structure on the covariance matrix of ε_i
- You need to assign products to categories before you estimate and you can't make mistakes!

Nested Logit

How does it actually look?

$$
\begin{aligned}
IV_{ig}(\theta) &= \log\left(\sum_{k\in G} \exp[x_k\beta/(1-\lambda_g)]\right) = E_\varepsilon[\max_{j\in G} u_{ij}] \\
s_{ij|g}(\theta) &= \frac{\exp[x_j\beta/(1-\lambda_g)]}{\sum_{k\in G}\exp[x_k\beta/(1-\lambda_g)]} \\
s_{ig}(\theta) &= \frac{\exp[IV_{ig}]^{1-\lambda_g}}{\sum_h \exp[IV_{ih}]^{1-\lambda_h}}
\end{aligned}
$$

- When $\lambda_g \to 0$ we get the IIA logit model (no correlation within nests)
- When $\lambda_g \to 1$ we get no across nest substitution.
- When $\lambda_g > 1$ we get something not necessarily consistent with utility maximization!

Nested Logit

How does it actually look?

$$\log\left(\frac{s_{ij|g}(\theta)}{s_{ik|g}(\theta)}\right) = (x_j - x_k)\cdot\frac{\beta}{1-\lambda_g}$$

- We are back to having the IIA property but now within the group G.
- We also have IIA across groups g, h
- λ_g and α govern the elasticities, which also have a block structure.
- Sometimes people refer to this as the product of two logits
- In the old days people used to estimate by fitting sequential IIA logit models – this is consistent but inefficient – you shouldn't do this today!
- Esimation happens via MLE. This can be tricky because the model is non-convex. It helps to substitute $\tilde{\beta} = \beta/(1-\lambda_g)$

Nested Logit

There are more potential generalizations though they are less frequently used:

- You can have multiple levels of nesting: first I select a size car (compact, mid-sized, full-sized) then I select a manufacturer, finally a car.

- You can have potentially overlapping nests: Yogurt brands are one nest, Yogurt flavors are a second nest. This way strawberry competes with strawberry and/or Dannon substitutes for Dannon.

Mixed Logit

We relax the IIA property by mixing over various logits:

$$
\begin{aligned}
u_{ijt} &= x_j\beta + \mu_{ij} + \varepsilon_{ij} \\
s_{ij} &= \int \frac{\exp[x_j\beta + \mu_{ij}]}{1 + \sum_k \exp[x_k\beta + \mu_{ik}]} f(\mu_i|\theta)
\end{aligned}
$$

- Each individual draws a vector μ_i of μ_{ij} (separately from ε).
- Conditional on μ_i each person follows an IIA logit model.
- However we integrate (or mix) over many such individuals giving us a mixed logit or heirarchical model (if you are a statistician)
- In practice these are not that different from linear random effects models you have learned about previously.
- It helps to think about fixing μ_i first and then integrating out over ε_i

Mixed/ Random Coefficients Logit

As an alternative, we could have specified an error components structure on ε_i.

$$U_{ij} = \beta x_{ij} + \underbrace{\nu_i z_{ij} + \varepsilon_{ij}}_{\tilde{\varepsilon}_{ij}}$$

- The key is that ν_i is unobserved and mean zero. But that x_{ij}, z_{ij} are observed per usual and ε_{ij} is IID Type I EV.
- This allows for a heteroskedastic structure on ε_i, but only one which we can project down onto the space of z.

An alternative is to allow for individuals to have random variation in β_i:

$$U_{ij} = \beta_i x_{ij} + \varepsilon_{ij}$$

Which is the random coefficients formulation (these are the same model).

Mixed/ Random Coefficients Logit

- Kinds of heterogeneity
  - We can allow for there to be two types of β_i in the population (high-type, low-type). latent class model.
  - We can allow β_i to follow an independent normal distribution for each component of x_{ij} such as $\beta_i = \overline{\beta} + \nu_i \sigma$.
  - We can allow for correlated normal draws using the Cholesky root of the covariance matrix.
  - Can allow for non-normal distributions too (lognormal, exponential). Why is normal so easy?
- The structure is extremely flexible but at a cost.
- We generally must perform the integration numerically.
- High-dimensional numerical integration is difficult. In fact, integration in dimension 8 or higher makes me very nervous.
- We need to be parsimonious in how many variables have unobservable heterogeneity.
- Again observed heterogeneity does not make life difficult so the more of that the better!

Mixed Logit

How does it work?

- Well we are mixing over individuals who conditional on β_i or μ_i follow logit substitution patterns, however they may differ wildly in their s_{ij} and hence their substitution patterns.
- For example if we are buying cameras: I may care a lot about price, you may care a lot about megapixels, and someone else may care mostly about zoom.
- The basic idea is that we need to explain the heteroskedasticity of $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$ what random coefficients do is let us use a basis from our X's.
- If our X's are able to span the space effectively, then an RC logit model can approximate any arbitrary RUM (McFadden and Train 2002).
- Of course if you have 1000 products and two random coefficients, you are asking for a lot.

Mixed/ Random Coefficients Logit

Suppose there is only one random coefficient, and the others are fixed:

- $f(\beta_i \theta) \sim N(\overline{\beta}, \sigma)$.
- We can re-write this as the integral over a transformed standard normal density

$$P_{ij}(\theta) = \int \frac{e^{V_{ij}(\nu_i, \theta)}}{\sum_k e^{V_{ik}(\nu_i, \theta)}} f(\nu_i) \partial \nu$$

- Monte Carlo Integration: Independent Normal Case
  - Draw ν_i from the standard normal distribution.
  - Now we can rewrite $\beta_i = \overline{\beta} + \nu_i \sigma$
  - For each β_i calculate $P_{ij}(\beta_i)$.
  - $\frac{1}{S} \sum_{s=1}^{S} P_{ij} = \widehat{P}_j^s$
- Gaussian Quadrature
  - Or we can draw a non-random set of points ν_i and corresponding weights w_i and approximate the integral to a high level of polynomial accuracy.

Quadrature in higher dimensions

- Quadrature is great in low dimensions – but scales badly in high dimensions.
- If we need $\mathrm{N_a}$ points to accurately approximate the integral in $\mathrm{d} = 1$ then we need $\mathrm{N_a^d}$ points in dimension d (using the tensor product of quadrature rules).
- There is some research on quadrature rules that nest and also how to carefully eliminate points so that the number doesn't grow so quickly.
- Try `sparse-grids.de`

Estimation

How do we actually estimate these models?

- In practice we should be able to do MLE.

$$\max_{\theta} \sum_{i=1}^{N} y_{ij} \log P_{ij}(\theta)$$

- When we are doing IIA logit, this problem is globally convex and is easy to estimate using Newton's Method.
- When doing nested logit or random coefficients logit, it generally is non-convex which can make life difficult.
- The tough part is generally working out what $\frac{\partial \log P_{ij}}{\partial \theta}$ is, especially when we need to simulate to obtain P_{ij}.
- It turns out that MSLE actually has consistent problems for fixed S. Why?
- Alternative? MSM/MoM type estimators (next time).

Mixed Logit: Estimation

- Just like before, we do MLE
- One wrinkle–how do we compute the integral?

$$
\begin{aligned}
s_{ij} &= \int \frac{\exp[x_j \beta_i]}{1 + \sum_k \exp[x_k \beta_i]} f(\beta_i | \theta) \\
&= \sum_{s=1}^{ns} w_s \frac{\exp[x_j(\overline{\beta} + \Sigma \nu_{is})]}{1 + \sum_k \exp[x_k(\overline{\beta} + \Sigma \nu_{is})]}
\end{aligned}
$$

- Option 1: Monte Carlo integration. Draw $NS = 1000$ or so samples of ν_i from the standard normal and set $w_i = \frac{1}{NS}$.
- Option 2: Quadrature. Choose ν_i and w_i according to a Gaussian quadrature rule. Like `quad` in MATLAB.
- Personally I get nervous about integrals in dimension greater than 5. People routinely have 20 or more though.

Mixed Logit: Hints

How bad is the simulation error?

- Depends how small your shares are.
- Since you care about $\log s_{jt}$ when shares are small, tiny errors can be enormous.
- Often it is pretty bad.
- I recommend sticking with quadrature at a high level of precision.
- `sparse-grids.de` provide efficient high dimensional quadrature rules.

Even More Flexibility (Fox, Kim, Ryan, Bajari)

Suppose we wanted to nonparametrically estimate $f(\beta_i|\theta)$ instead of assuming that it is normal or log-normal.

$$s_{ij} \quad = \quad \int \frac{\exp[x_j \beta_i]}{1 + \sum_k \exp[x_k \beta_i]} f(\beta_i|\theta)$$

- Choose a distribution $g(\beta_i)$ that is more spread out that $f(\beta_i|\theta)$
- Draw several β_s from that distribution (maybe 500-1000).
- Compute $\hat{s}_{ij}(\beta_s)$ for each draw of β_s and each j.
- Holding $\hat{s}_{ij}(\beta_s)$ fixed, look for w_s that solve

$$\min_w \left(s_j - \sum_{s=1}^{ns} w_s \hat{s}_{ij}(\beta_s) \right)^2 \quad \text{s.t.} \sum_{s=1}^{ns} w_s = 1, \quad w_s \geq 0 \quad \forall s$$

Even More Flexibility (Fox, Kim, Ryan, Bajari)

- Like other semi-/non- parametric estimators, when it works it is both general and very easy.
- We are solving a least squares problem with constraints: positive coefficients, coefficients sum to 1.
- It tends to produce sparse models with only a small number of β_s getting positive weights.
- This is way easier than solving a random coefficients logit model with all but the simplest distributions.
- There is a bias-variance tradeoff in choosing $g(\beta_i)$.
- Incorporating parameters that are not random coefficients loses some of the simplicity.
- I have no idea how to do this with large numbers of fixed effects.

Convexity and Computation

Convexity

An optimization problem is convex if

$$\min_{x} f(x) \quad \text{s.t.} \quad h(x) \leq 0 \quad Ax = 0$$

- $f(x), h(x)$ are convex (PSD second derivative matrix)
- Equality Constraint is affine

Some helpful identities about convexity

- Compositions and sums of convex functions are convex.
- Norms $||$ are convex, $\max$ is convex, $\log$ is convex
- $\log(\sum_{i=1}^{n} \exp(x_i))$ is convex.
- Fixed Points can introduce non-convexities.
- Globally convex problems have a unique optimum

Properties of Convex Optimization

- If a program is globally convex then it has a unique minimizer that will be found by convex optimizers.
- If a program is not globally convex, but is convex over a region of the parameter space, then most convex optimization routines find any local minima in the convex hull
- Convex optimization routines are unlikely to find local minima (including the global minimum) if they do not begin in the same convex hull as the optimum (starting values matter!).
- Most good commercial routines are clever about dealing with multiple starting values and handling problems that are well approximated by convex functions.
- Good Routines use information about sparseness of Hessian – this generally determines speed.

Nested Logit Model

FIML Nested Logit Model is Non-Convex

$$\min_{\theta} \sum_{j} q_j \ln P_j(\theta) \quad \text{s.t.} \quad P_j(\theta) = \frac{e^{x_j\beta/\lambda} \left(\sum_{k\in g_l} e^{x_j\beta/\lambda}\right)^{\lambda-1}}{\sum_{\forall l'} \left(\sum_{k\in g_l'} e^{x_j\beta/\lambda}\right)^{\lambda}}$$

This is a pain to show but the problem is with the cross term $\frac{\partial^2 P_j}{\partial\beta\partial\lambda}$ because $\exp[x_j\beta/\lambda]$ is not convex.

A Simple Substitution Saves the Day: let $\gamma = \beta/\lambda$

$$\min_{\theta} \sum_{j} q_j \ln P_j(\theta) \quad \text{s.t.} \quad P_j(\theta) = \frac{e^{x_j\gamma} \left(\sum_{k\in g_l} e^{x_j\gamma}\right)^{\lambda-1}}{\sum_{\forall l'} \left(\sum_{k\in g_l'} e^{x_j\gamma}\right)^{\lambda}}$$

This is much better behaved and easier to optimize.

Nested Logit Model

	Original[1]	**Substitution**[2]	**No Derivatives**[3]
Parameters	49	49	49
Nonlinear λ	5	5	5
Likelihood	2.279448	2.279448	2.27972
Iterations	197	146	352
Time	59.0 s	10.7 s	192s

Discuss Nelder-Meade

Computing Derivatives

A key aspect of any optimization problem is going to be computing the derivatives (first and second) of the model. There are some different approaches

- Numerical: Often inaccurate and error prone (why?)
- Pencil and Paper: this tends to be mistake prone – but often actually the fastest
- Automatic (AMPL): Software brute forces through a chain rule calculation at every step (limited language).
- Symbolic (Maple/Mathematica): software "knows" derivatives of certain objects and can do its own simplification. (limited language).