Part 7: Multinomial Choice

Chris Conlon

Microeconometrics

September 12, 2016

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.).

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.

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 $(t = 1)$:

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

Setup

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

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

Setup

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

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

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

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

Setup

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

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

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
- 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: P_{ij} is everywhere continuously differentiable in V_{ij}.
- Bound $P_{ij} \in (0, 1)$ 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 characterstics $\theta p_j / Y_i$.

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!).

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).

Logit

- Logit has closed form choice probabilities

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

- 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

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

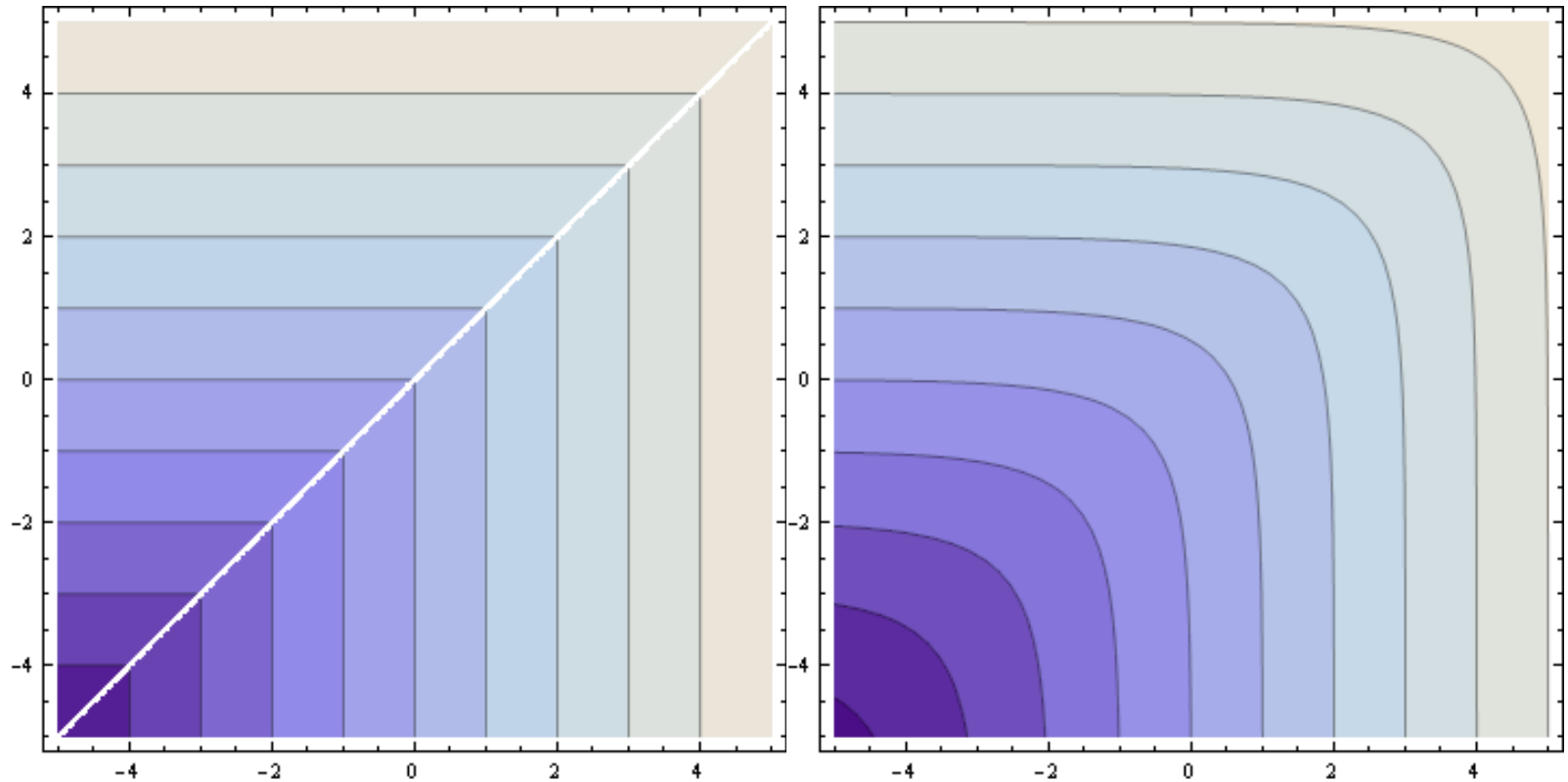
- Expected utility of best option (without knowledge of realized ε_i) does not depend on ϵ_{ij}.
- This is a globally concave function in V_{ij} (more on that later).
- Allows simple computation of ΔCS for consumer welfare.

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



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 $Var(\varepsilon^*/\sigma) = Var(\varepsilon) = \pi^2/6$

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

- 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 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

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

- 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. $P_c = P_{bb} = \frac{1}{2}$ so that $\frac{P_c}{P_{bb}} = 1$.
- Now we introduce a red bus rb that is identical to bb. Then $\frac{P_{rb}}{P_{bb}} = 1$ and $P_c = P_{bb} = P_{rb} = \frac{1}{3}$ 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 $P_c = \frac{1}{2}$ and $P_{bb} = P_{rb} = \frac{1}{4}$.
- 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 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 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 $\hat{\beta}$, 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

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

And Elasticity:

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

With cross effects:

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

And Elasticity:

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

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

Proportional Substitution

Cross elasticity doesn't really depend on j.

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

- 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 $\tilde{P}_{ij} = \frac{P_{ij}}{1-P_{ik}}$ for all other j.
- This might be a desirable property but probably not.

Alternatives to IIA

IIA doesn't seem like a particularly desireable property. How can we relax it?

- The problem arises because we required that ε_{ij} were IID.
- We would like to allow for a more general heteroskedastic structure on ε_{ij}
- Options
  - Lots of observable individual characteristics and interact them with the x_j's. Then maybe IIA is mostly about smoothness not about undesireable properties.
  - Multivariate Probit allows for $\varepsilon_i \sim N(0, \Omega)$.
  - Put some more structure on the problem: Assume that ε_i has a block structure. Assign choices to categories and allow for more correlation between choices in prespecified categories.
  - Mixed logit allows for unobserved heterogeneity ν_i that we can interact with x_j from some arbitrary distribution $f(\nu_i|\theta)$. Conditional on a ν_i individual behavior is still logit. This is actually a basis structure on ε_i.

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

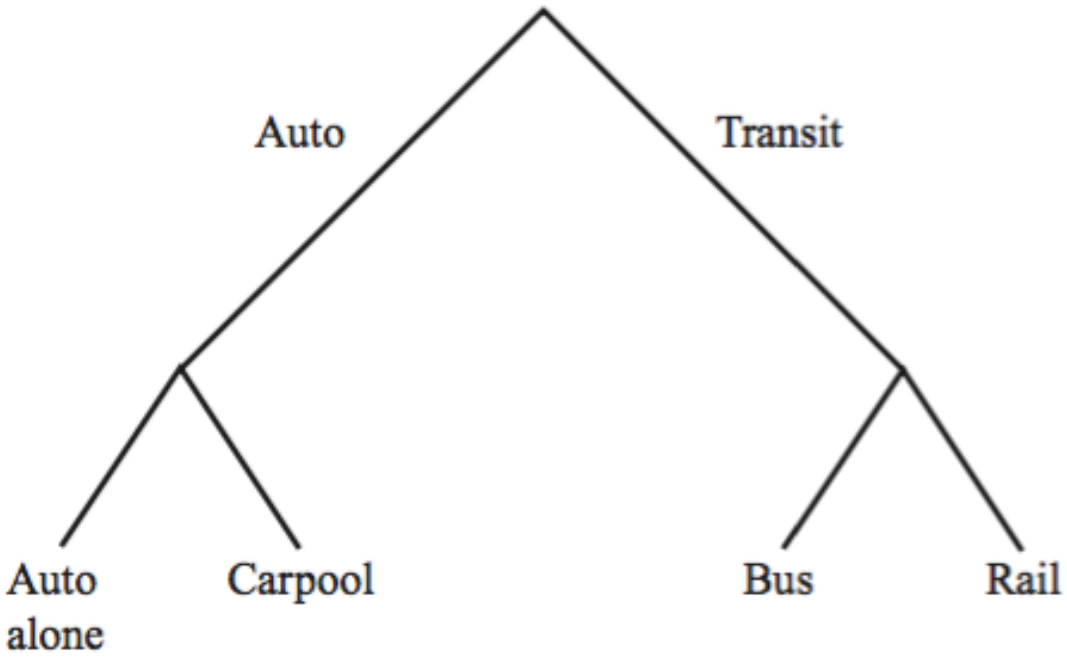


Figure 4.1. Tree diagram for mode choice.

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 $\lambda_g \cdot \underbrace{\log\left(\sum_{k\in J_g} e^{V_{ik}}\right)}_{IV_{ig}}$.

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

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.

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

For each individual i, the resulting choice probability follows a logit:

$$P_{ij} = \int \frac{e^{V_{ij}(\beta_i)}}{\sum_k e^{V_{ik}(\beta_i)}} f(\beta_i|\theta)\partial\beta$$

This structure is quite general:

- The choice probabilities are know a function of unknown parameters θ.
- 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?

Mixed/ Random Coefficients Logit

- 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/ Random Coefficients Logit

How do we approximate:

$$P_{ij} = \int \frac{e^{V_{ij}(\beta_i)}}{\sum_k e^{V_{ik}(\beta_i)}} f(\beta_i|\theta)\partial\beta$$

- Monte Carlo Integration
  - Draw β_i from the candidate distribution. $[\beta_i^{(1)}, \beta_i^{(2)}, \ldots \beta_i^{(s)}]|\theta$.
  - For each β_i calculate $P_{ij}(\beta_i)$.
  - $\frac{1}{S}\sum_{s=1}^{S} P_{ij} = \widehat{P_j^s}$

The way we usually get correlated normal variables (or any normal variables) is to transform independent normals appropriately.

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 N_a points to accurately approximate the integral in $d = 1$ then we need 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).

Semi-parametric Alternative

Fox Kim Bajari Ryan (QE 2011) propose a nice alternative:

$$g_j(x_i, \beta^r) = \frac{\exp(x_{ij}\beta^r)}{1 + \sum_{j=1}^{J} \exp(x_{ij}\beta^r)}$$

$$\theta^r \geq 0 \sum_{r=1}^{R} \theta^r = 1$$

$$E[Y_{ij} - \sum_{r=1}^{R} \theta^r g_j(x_i, \beta^r)] = 0$$

Via constrained LLS.

Convexity

An optimization problem is convex if

$$\min_x f(\mathbf{x}) \quad s.t. \quad h(\mathbf{x}) \leq 0 \quad A\mathbf{x} = 0$$

- $f(\mathbf{x}), h(\mathbf{x})$ 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).