

# MULTINOMIAL DISCRETE CHOICE

CHRIS CONLON

APPLIED ECONOMETRICS II

MARCH 15, 2019

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  $P_{ij t}$ .
- 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 ( $t = 1$ ):

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

# NONPARAMETRIC SETUP

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

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

# NONPARAMETRIC SETUP

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

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

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

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

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

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  $\varepsilon_{ij}$  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

$$U_b = X_b\beta + k_b + \varepsilon_b$$

$$U_c = X_c\beta + k_c + \varepsilon_c$$

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

# 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  $\Omega$  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  $\lambda$  factor doesn't change any statements about  $U_{ij} > U_{ik}$ .
- We normalize this by fixing the variance of  $\varepsilon_{ij}$  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

# OBSERVED HETEROSKEDASTICITY

Consider the case where  $Var(\varepsilon_{ij}^B) = \sigma^2$  and  $Var(\varepsilon_{ij}^C) = k^2\sigma^2$  :

- We can estimate

$$U_{ij} = x_j\beta + \varepsilon_{ij}^B$$

$$U_{ij} = x_j\beta + \varepsilon_{ij}^C$$

becomes:

$$U_{ij} = x_j\beta + \varepsilon_{ij}$$

$$U_{ij} = x_j\beta/k + \varepsilon_{ij}$$

- Some interpret this as saying that in segment  $C$  the unobserved factors are  $\hat{k}$  times larger.

## 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 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 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  $\varepsilon_i$ ) does not depend on  $\varepsilon_{ij}$ .
- This is a globally concave function in  $V_{ij}$  (more on that later).
- Allows simple computation of  $\Delta CS$  for consumer welfare.

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

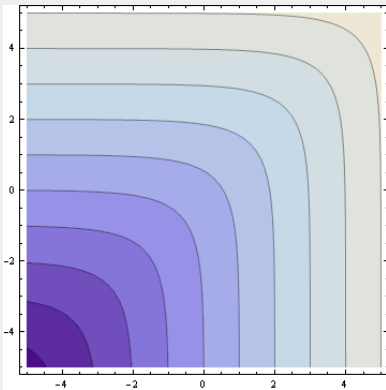
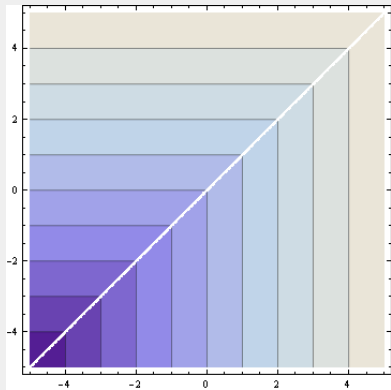
# ALTERNATIVE INTERPRETATION

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

$$\log \frac{s_{ij}(\theta)}{s_{ik}(\theta)} = \mathbf{x}_{ij}\beta_j - \mathbf{x}_{ik}\beta_k \rightarrow \mathbf{x}_i \cdot (\beta_j - \beta_k)$$

- We only identify the **difference in indirect utilities** not the levels.
- This is a feature and not a bug. Why?

# MULTINOMIAL LOGIT: IDENTIFICATION

As another idea suppose we add a constant  $C$  to each  $\beta_j$ .

$$s_{ij} = \frac{\exp[\mathbf{x}_i(\beta_j + C)]}{\sum_k \exp[\mathbf{x}_i(\beta_k + C)]} = \frac{\exp[\mathbf{x}_i C] \exp[\mathbf{x}_i \beta_j]}{\exp[\mathbf{x}_i C] \sum_k \exp[\mathbf{x}_i \beta_k]}$$

- 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 what?

# MULTINOMIAL LOGIT: IDENTIFICATION

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

$$s_{ij} = \frac{\exp[\mathbf{x}_i \beta_j]}{1 + \sum_k \exp[\mathbf{x}_i \beta_k]}$$

- 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  $Var(\varepsilon^*) = \sigma^2\pi^2/6$ .
- Without changing behavior we can divide by  $\sigma$  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  $\beta$  is rescaled by  $\sigma$ . This implies that only the ratio  $\beta^*/\sigma$  is identified.
- Coefficients are relative to variance of unobserved factors. More unobserved variance  $\rightarrow$  smaller  $\beta$ .
- Ratio  $\beta_1/\beta_2$  is invariant to the scale parameter  $\sigma$ .

# TASTE VARIATION

- Logit allows for taste variation across individuals if two conditions are met: **individual level data** and **interact observed characteristics** only.
- We often want to allow for something like
$$U_{ij} = x_j\beta_i - \alpha_ip_j + \varepsilon_{ij}.$$
- 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  $\varepsilon_{ij}$ .

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

- 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 satisfied 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  $\beta$  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.

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

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:

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

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  $j$ .

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

- 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 might be a desirable property but probably not.

# MULTINOMIAL LOGIT: ESTIMATION WITH INDIVIDUAL DATA

Estimation is straightforward via Maximum Likelihood (MLE):

$$L(\mathbf{y}|\mathbf{x}, \theta) = \prod_{i=1}^N \frac{n_i!}{\underbrace{\prod_{j=1}^J y_{ij}!}_{C(\mathbf{y})}} \prod_{j=1}^J s_{ij}(x_{ij}, \theta)^{y_{ij}}$$

$$ll(\mathbf{y}|\mathbf{x}, \theta) = \sum_{i=1}^N \log(C(\mathbf{y})) + \sum_{i=1}^N \sum_{j=1}^J y_{ij} \log(s_{ij}(x_{ij}, \theta))$$

$$l(\mathbf{y}|\mathbf{x}, \theta) \approx \sum_{i=1}^N \sum_{j=1}^J y_{ij} \log(s_{ij}(x_{ij}, \theta))$$

- We can ignore the combinatorial term (with the factorials) since it does not affect the location of the maximum (it is additive and doesn't depend on  $\theta$ ).

# MULTINOMIAL LOGIT: INCLUSIVE VALUE

To be more specific:

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

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

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

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

- 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  $s_j(\mathbf{x}_i, \theta)$ . When it doesn't change across observations: we can aggregate.
- It functions as if we have a representative consumer up to  $\varepsilon_i$ .
- We can use this idea to go from individual level to market demand:  $q_j(\mathbf{x}_i) = N_i s_{ij}(\theta)$ .

# MULTINOMIAL LOGIT: ELASTICITY

An important output from a demand system are elasticities

- An important element in  $\mathbf{x}_i$  are prices  $[p_1, \dots, p_J]$
- Helpful to write  $u_{ij} = x_j\beta - \alpha p_j$  (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! (Why is this a bad idea?)
- $\eta_{jk} = \frac{\partial q_j}{\partial p_k} \frac{p_k}{q_j} = -\alpha p_k s_k$ .
- The cross price elasticity doesn't depend on which product  $j$  you are talking about!



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  $\Sigma$  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  $\rho_{12}$  is large if 1, 2 are similar products.
- Much more flexible than Logit

## Downside

- $\Sigma$  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  $\varepsilon_{ij}$  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  $\varepsilon_{ij}$  (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**.

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

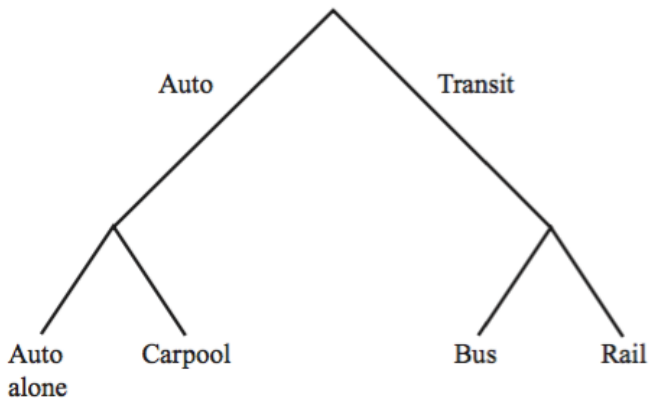


Figure 4.1. Tree diagram for mode choice.

# 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  $\varepsilon_{ij}$ .

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

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

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)
- $\lambda_g \rightarrow 0$  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).
- $\lambda$  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.

# 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  $\varepsilon_i$
- You need to assign products to categories **before you estimate** and you can't make mistakes!

How does it actually look?

$$IV_{ig}(\theta) = \log \left( \sum_{k \in G} \exp[x_k \beta / (1 - \lambda_g)] \right) = E_\epsilon [\max_{j \in G} u_{ij}]$$

$$s_{ij|g}(\theta) = \frac{\exp[x_j \beta / (1 - \lambda_g)]}{\sum_{k \in G} \exp[x_k \beta / (1 - \lambda_g)]}$$

$$s_{ig}(\theta) = \frac{\exp[IV_{ig}]^{1-\lambda_g}}{\sum_h \exp[IV_{ih}]^{1-\lambda_h}}$$

# 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  $g, h$
- $\lambda_g$  and  $\alpha$  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!
- Estimation happens via MLE. This can be tricky because the model is non-convex. It helps to substitute  $\tilde{\beta} = \beta / (1 - \lambda_g)$

# PARAMETRIC IDENTIFICATION

Look at derivatives:

$$\frac{\partial s_{j|g}}{\partial X_j} = \beta_1 s_{j|g}(1 - s_{j|g})$$

$$\frac{\partial s_g}{\partial X} = (1 - \lambda)\beta_1 s_g(1 - s_g)$$

$$\frac{\partial s_g}{\partial J} = \frac{1 - \lambda}{J} s_g(1 - s_g)$$

- We get  $\beta$  by changing  $x_j$  within group
- We get nesting parameter  $\lambda$  by varying  $X$
- We don't have any parameters left to explain changing number of products  $J$ .

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  $\mu_i$  of  $\mu_{ij}$  (separately from  $\varepsilon$ ).
- Conditional on  $\mu_i$  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  $\mu_i$  first and then integrating out over  $\varepsilon_i$

## MIXED/ RANDOM COEFFICIENTS LOGIT

As an alternative, we could have specified an error components structure on  $\varepsilon_i$ .

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

- The key is that  $\nu_i$  is unobserved and mean zero. But that  $x_{ij}, z_{ij}$  are observed per usual and  $\varepsilon_{ij}$  is IID Type I EV.
- This allows for a heteroskedastic structure on  $\varepsilon_i$ , 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  $\beta_i$ :

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

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  $\beta_i$  in the population (high-type, low-type). **latent class model**.
  - ▶ We can allow  $\beta_i$  to follow an independent normal distribution for each component of  $x_{ij}$  such as  $\beta_i = \bar{\beta} + \nu_i \sigma$ .
  - ▶ 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  $\beta_i$  or  $\mu_i$  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  $Cov(\varepsilon_i, \varepsilon_j)$  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:

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

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

- Monte Carlo Integration: Independent Normal Case

- ▶ Draw  $\nu_i$  from the standard normal distribution.
- ▶ Now we can rewrite  $\beta_i = \bar{\beta} + \nu_i\sigma$
- ▶ For each  $\beta_i$  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  $\nu_i$  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).

# MIXED LOGIT: ESTIMATION

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

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

- Option 1: Monte Carlo integration. Draw  $NS = 1000$  or so samples of  $\nu_i$  from the standard normal and set  $w_i = \frac{1}{NS}$ .
- Option 2: Quadrature. Choose  $\nu_i$  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.

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.

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

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

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



## 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  $\beta_s$  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

## 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  $\| \cdot \|$  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

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

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  $\gamma = \beta / \lambda$

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

# NESTED LOGIT MODEL

	<b>Original<sup>1</sup></b>	<b>Substitution<sup>2</sup></b>	<b>No Derivatives<sup>3</sup></b>
Parameters	49	49	49
Nonlinear $\lambda$	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).