Preference Heterogeneity and Mixed Logit

Meredith Fowlie
UC Berkeley

Spring 2021

Outline for today

- Introducing the Berry Transformation .
- Introducing Mixed Logit

Introducing the 'Berry Transformation'

The assigned discussion paper uses a standard trick to implement familiar IV methods in this non-linear demand system context. Page 19 summarizes the approach as follows:

> *This procedure is common to address endogeneity in the estimation of random coefficients logit models (Berry, Levinsohn and Pakes, 1995, 2004), of which the nested logit is a simple case. The key idea is to invert market shares to solve for mean indirect utilities, which then allows for linear IV estimates in the second stage that are unbiased despite the endogeneity of price to quality (Berry, 1994).*

Let's break this down...

The (Steve) Berry transformation

The first step formalizes the relationship between consumer-level choice probabilities and market shares. To simplify notation, let $\delta_j = X_j'\beta + \alpha p_j$.

$$\begin{aligned} U_{njt} &= \delta_{jt} + \varepsilon_{njt} \\ \Pr(U_{nit} &\geq U_{njt} \forall j) = \frac{\exp(\delta_{it})}{1 + \sum_j \exp(\delta_{jt})}, \end{aligned}$$

$1 = e^0$ = outside option utility normalized to 0

In this simple CL framework, the estimated market share of product j is simply the predicted choice probability:

$$\widehat{s}_{it}(\delta_{it}) = \Pr(U_{ni} \geq U_{nj} \forall j) = \frac{\exp(\delta_i)}{1 + \sum_j \exp(\delta_j)}. \quad (1)$$

To define market shares, we assume a choose nothing option with latent utility normalized to 0.

System of share equations:

Next step: define market share equations $\widehat{s}_{it}(\delta_{it})$:

$$\widehat{s}_0 = \frac{1}{1 + \sum_j \exp(\delta_j)} \tag{2}$$

$$\widehat{s}_1 = \frac{\exp(\delta_1)}{1 + \sum_j \exp(\delta_j)} \tag{3}$$

$$\vdots \tag{4}$$

$$\widehat{s}_J = \frac{\exp(\delta_J)}{1 + \sum_j \exp(\delta_j)} \tag{5}$$

With these share equations, we can in principle solve for the δ_{jt} that fit/match the observed shares exactly.

Reformulating...

In the CL model, this can be done analytically. Taking logs:

$$\log \widehat{s}_0 = -\log(1 + \sum_j \exp(\delta_j)) \tag{6}$$

$$\log \widehat{s}_1 = \delta_1 - \log(1 + \sum_j \exp(\delta_j)) \tag{7}$$

$$\vdots \tag{8}$$

$$\log \widehat{s}_J = \delta_J - \log(1 + \sum_j \exp(\delta_j)) \tag{9}$$

This yields:

diff of log(share of consumers choosing j) and log(share of consumers not choosing any options)

$$\log \widehat{s}_j - \log \widehat{s}_0 = \delta_j = X_j'\beta + \alpha p_j + \xi_j \tag{10}$$

The key intuition: There exists a unique $\vec{\delta}$ such that the predicted shares equal actual (observed) shares. This means that we can write $\delta(\theta_2)$.

Limitations of conditional logit

We now have a system of * linear *share equations that are implied by the conditional logit model. Very convenient. We have just massaged this non-linear demand system into a form that we can estimate using *linear* regression!!

$$\log \widehat{s}_j - \log \widehat{s}_0 = \delta_j = X_j'\beta + \alpha p_j + \xi_j \tag{11}$$

Concerns?

Limitations of conditional logit

We now have a system of * linear *share equations that are implied by the conditional logit model. Very convenient. We have just massaged this non-linear demand system into a form that we can estimate using *linear* regression!!

$$\log \widehat{s}_j - \log \widehat{s}_0 = \delta_j = X_j'\beta + \alpha p_j + \xi_j \tag{11}$$

Concerns? Same issues we have been discussing. Note that:

$$\begin{aligned} \frac{\partial s_{it}}{\partial p_{jt}} &= \alpha \left(\frac{\exp(\delta_i)}{1 + \sum_j \exp(\delta_j)} \right) \left(\frac{\exp(\delta_j)}{1 + \sum_J \exp(\delta_j)} \right) \\ &= \alpha s_i s_j \end{aligned}$$

The cross-price elasticity of product j with respect to the price of k depends only on the features of these two products (IIA).

The Beauty of the Berry transform?

We now have a system of * linear *share equations that are implied by the conditional logit model. Very convenient. We have just massaged this non-linear demand system into a form that we can estimate using *linear* regression!!

$$\log \widehat{s}_j - \log \widehat{s}_0 = \delta_j = X_j'\beta + \alpha p_j + \xi_j \tag{12}$$

The key intuition: There exists a unique $\beta' X$ such that the predicted shares equal actual (observed) shares.

A key advantage?: Inverting market shares to solve for mean indirect utilities allows us to implement our standard IV tools and tricks.

Outline for today

- Introducing the Berry Transformation .
- Introducing Mixed Logit

Reasons to love the mixed logit framework

1. Extremely general! Researcher can choose the distribution that best fits the empirical setting.
2. Intuitively accommodates random taste variation across people.
3. Accommodates correlation in unobserved attributes over choices/time.
4. No more IIA!

Motivating the mixed logit

Many ways to motivate the mixed logit specification.

The most common (IO, marketing, EEE) and most intuitive (for us economists) starts from the same RUM (or RCM) point of departure.

$$\begin{aligned} U_{ni} &= \beta_n X_{ni} + \varepsilon_{ni} \\ & \quad \varepsilon_{ni} \tilde{} \; iid \text{ EV} \\ & \quad \beta_n \tilde{} \; f(\beta|\theta) \end{aligned}$$

The small but important difference? β can vary randomly across agents. This reflects the fact that different decision-makers have different tastes/preferences.

A small but important difference between ML and CL

- We no longer assume that these taste paramteters are fixed across agents.
- The f(θ) describes the density of these taste coefficientts, where θ is a vector containing the parameters of the distribution of taste parameters.
- For example, if the β is normally distributed, θ contains the mean and variance.
- We now have two types of parameters:
  1. The β_n vary across agents (heterogeneous tastes).
  2. The parameters θ define the assumed distribution in the population.

Conditional choice probabilities

- If we knew the β_n, we could condition on it and get back to the tractable CL choice probability.
- This conditional probability is just the standard logit evaluated at β_n:

$$P_{ni}(\beta_n) = \frac{\exp(\beta_n' X_{ni})}{\sum_j \exp(\beta_n' X_{nj})}$$

- This is a nice closed form .. no simulation needed!
- BUT we don't know β_n.. so we can't really condition on it..so what do we do?

Unconditional choice probabilities

To obtain the unconditional choice probability, we need to integrate over the density of β :

$$P_{ni} = \int \frac{\exp(\beta_n' X_{ni})}{\sum_j \exp(\beta_n' X_{nj})} f(\beta) d\beta$$

- This is essentially a weighted average of the logit probabilities evaluated at different β values.
- The weights are given by the density $f(\beta)$.
- This is sometimes called the "mixing" distribution as it defines the weights in this mix of alternative logit functions.
- In most applications, this mixing function is continuous, although there are cases where you might want β to take on a discrete set of values (e.g. latent class models).

Convenient error partitioning

$$
\begin{aligned}
U_{ni} &= \beta_n X_{ni} + \varepsilon_{ni} \\
&\quad \varepsilon_{ni} \tilde{} \; iid \text{ EV} \\
&\quad \beta_n \tilde{} \; f(\beta|\theta)
\end{aligned}
$$

- We continue to assume that the ε are distributed iid EV. So part of the integration can be done analytically.
- This model can be generalized to accommodate both observable and unobservable taste variation.
- An alternative motivation places more emphasis on capturing substitution patterns across choices parsimoniously and realistically.

How do you estimate this thing?

$$
\begin{aligned}
U_{ni} &= \beta_n X_{ni} + \varepsilon_{ni} \\
&\quad \varepsilon_{ni} \tilde{} \; iid \text{ EV} \\
&\quad \beta_n \tilde{} \; f(\beta|\theta)
\end{aligned}
$$

How do you estimate this thing?

$$
\begin{aligned}
U_{ni} &= \beta_n X_{ni} + \varepsilon_{ni} \\
&\varepsilon_{ni} \tilde{} \; iid \text{ EV} \\
&\beta_n \tilde{} \; f(\beta|\theta)
\end{aligned}
$$

- Integrate explicitly over ε_{in} given β_n.
- Integrate numerically (via simulation) over the assumed density $f(\beta_n|\theta)$
- Mechnically, how can we evaluate the integral of a statistic over the assumed density?

Simulation!

If we have a well defined distribution, we can randomly draw from it...

1. Draw R times from the assumed distribution $f(\beta_n|\theta)$.
2. Construct R values of the conditional choice probability evaluated using these R values.
3. Take an average. This gives you your simulated choice probability:

$$P_{ni} = \frac{1}{R}\sum_r \frac{\exp(\beta'_{nr}X_{ni})}{\sum_j \exp(\beta'_{nr}X_{nj})}$$

Insert these simulated probabilities into your simulated log-likelihood function:

$$SLL(\theta) = \sum_n \sum_j y_{nj} \ln\left(\frac{1}{R}\sum_r \frac{\exp(\beta'_{nr}X_{ni})}{\sum_j \exp(\beta'_{nr}X_{nj})}\right)$$

Your ML parameter estimates of θ are those that maximize the value of this simulated likelihood function.

Why is this referred to as convenient error partitioning?

- Partition the error into a component that is distributed in a way that yields choice probabilities that are analytically integrable AND an error that is distributed in a way that makes sense given the choice context.
- Note that an integral over a density is essentially a weighted average.
- We can approximate this with a simulated weighted average.

Distributional assumptions?

- What you assume about the distribution of the random components really depends on the context.
- Suppose you assume the coefficient is distributed randomly in the population : $\beta \tilde{} N(b, s^2)$
- To simulate drawing from this distribution, take a random draw from a standard normal. Multiply by s. Add b. You are done!
- When would you want to choose a distribution *other* than a normal?
- Price coefficient? Log normal a better choice (price enters negatively). Draws from a log normal: $\beta = exp(b + s\eta)$.

Mixed logit to the rescue?

The mixed logit (aka random coefficients logit, error components model) addresses the three limitations we've been discussing.

- Models/accounts for random taste variation/heterogeneity.
- Can accommodate orrelation in unobserved factors across choices/time.
- Substitution patterns uncovered versus imposed.

The key difference: ML model is not tied to any particular distributional assumptions for the unobserved component. Researcher can choose the distribution that best fits the empirical setting.

Panel data

Suppose we observe T_n choice situations/outcomes for person n. Note this number can vary across people. So now we have:

$$\begin{aligned} U_{nti} &= \beta_n X_{nti} + \varepsilon_{nti} \\ & \varepsilon_{nti} \tilde{} \ EV1 \\ & \beta_n \tilde{} f(\beta|\theta) \end{aligned}$$

The outcome of these choice situations is a vector: $y_n = \{y_1 ... y_{T_n}\}$

Conditional on the β for that person, the probability of observing the sequence of decisions is:

$$\Pr ob(y|\beta) = \prod_{t=1}^{T_n} \frac{e^{\beta_n' x_{nty_{nt}}}}{\sum e^{\beta_n' x_{njt}}}$$

Given assumed distribution of ε, this is just the product of logit choice probabilities.

Panel data, unconditional probabilities

The unconditional probability integrates this over the density of β.

$$(y|\theta) = P(y|\beta)f(\beta|\theta)d\beta$$

How does this accommodate serial correlation?

Panel data, unconditional probabilities

The unconditional probability integrates this over the density of β.

$$(y|\theta) = P(y|\beta)f(\beta|\theta)d\beta$$

How does this accommodate serial correlation?

$$\begin{aligned} U_{nti} &= \beta_n X_{nti} + \varepsilon_{nti} \\ &= bX_{nti} + s\eta_n X_{nit} + \varepsilon_{nit} \\ &= bX_{nit} + e_{nit} \end{aligned}$$

- Note that $cov(e_{nit}, e_{nit-1}) = s^2$ depends on the variance of the β and the extent of correlation of the X_{nit} across choices.
- Error component now correlated across choices made by the same agent.

More realistic substution patterns

- The restrictive assumptions of the CL model placed very strong restrictions on substitution patterns.
- Problem stems from the iid errors that are uncorrelated across choices and people.
- With heterogeneous preferences, the agent who chose a more fuel efficient car has a β vector that weights fuel efficiency more heavily than average.
- In other words, error components are correlated across choices with similar choice attributes.
- This generates stronger substitution between more similar choices.

Substitution patterns

Random coefficients/error components that are correlated across alternatives create correlation in random components across alternatives. To see this:

$$\begin{aligned} U_{ni} &= \beta_n X_{ni} + \varepsilon_{ni} \\ &= (b + \upsilon_n) X_{ni} + \varepsilon_{ni} \\ &= b X_{ni} + \upsilon_n X_{ni} + \varepsilon_{ni} \\ &= b X_{ni} + e_{ni} \end{aligned}$$

- If you look at the correlation in the errors across alternatives within a decision-maker, they now depend on the choice characteristics.
- The ratio of choice probabilities now depends on all of the data, including attributes of all choice alternatives.

What about those individual-specific parameters?

- Once you have estimated your ML model, you have in hand estimates of the parameters of the distributions of the β coefficients in the population (e.g. WTP for fuel efficiency).
- But you might want to know where a particular agent is within this distribution. (Why??)
- Each agent's choices reveal something about her preferences. We know how her choices differ from others in the population.
- How can we formalize inference from the population to an individual ?

Individual specific parameters

Distinguish between two distributions of the random taste parameters:

1. The distribution of the parameter in the population: $\beta \tilde{} f(\beta|\theta)$. This conditions only on θ.
2. The distribution $h(\beta|y, X, \theta)$ This is the distribution of β in the subpopulation who, when faced with the choice situation characterized by X would make the set of choices y.

The h( ) distribution depends on observed choices y, the choice set X and the population parameters θ.

Loosely speaking (some weighting issues aside), if we sum across all these h distributions, we should get the f (population) distribution back.

Individual taste parameters

$$\begin{aligned} U_{nit} &= \beta_n' X_{nit} + \epsilon_{nit} \\ & \quad \varepsilon_{nit} \tilde{} \, iid \text{ EV} \\ & \quad \beta_n \tilde{} \, f(\beta|\theta) in population \end{aligned}$$

Outcome vector is y_n. Conditional on persons tastes β_n:

$$P(y_n|\beta) = \prod_{t=1}^{T_n} \frac{e^{\beta_n' X_{nty_{nt}}}}{\sum_j e^{\beta_n' X_{njt}}}$$

Just a product of logit formulas!

The unconditional probability (probability of the choices given θ):

$$P(y_n|\theta) = \int P(y_n|\beta) f(\theta) d\beta$$

Bayes Theorem (not Bayesian estimation!)

What can we learn from agents' choices about the h distribution for each decision maker in our sample (realizing each decision maker belongs to a particular sub-population characterized by y and X).

Bayes rule (which relates conditional and marginal distributions) tells us that the joint probability of β and y can be expressed in two equivalent ways:

1. The probability of β given observed choices y times the probability of observed choices given θ.
2. The probality of the observed choices given β (a conditional distribution) times the probability of β (marginal distribution)

$$h(\beta|y,\theta)P(y|\theta) = P(y|\beta)f(\beta|\theta)$$

This gives us a way to calculate h!

Individual specific parameters

Rearranging this implication of Bayes rule:

$$h(\beta|y_n, X_n, \theta) = \frac{P(y_n|X_n, \beta)f(\beta|\theta)}{P(y_n|X_n, \theta)}$$

- $P(y_n|X_n, \beta)$ is the probability of making the observed choices y given β
- $P(y_n|X_n, \theta)$ is the probability of making the observed choices given the population parameters.
- $f(\beta|\theta)$ is the marginal distribution of β

Recall from the unconditional mixed logit choice probability:

$$P(y_n|X_n, \theta) = \int P(y_n|X_n, \beta)f(\beta|\theta)d\beta$$

So if we make this substitution we have:

$$h(\beta|y_n, X_n, \theta) = \frac{P(y_n|X_n, \beta)f(\beta|\theta)}{\int P(y_n|X_n, \beta)f(\beta|\theta)d\beta}$$

Individual specific parameters

$$h(\beta|y_n, X_n, \theta) = \frac{P(y_n|X_n, \beta)f(\beta|\theta)}{\int P(y_n|X_n, \beta)f(\beta|\theta)d\beta}$$

- Numerator: population density $f(\beta|\theta)$ weighted by the likelihood of a person's observed choices y for each possible value of β.
- Denominator: Integral of numerator (a normalizing constant).
- Intuitively - move through the distribution of β in the population and weight each value by the probability that this agent would have made the choices he made had he had that β.

Individual specific parameters

- This is not Bayesian estimation! This is an implication of your MLE of θ which you can derive/uncover with the help of the identity that is Bayes theorem.
- Cool... but why bother?