Discrete Choice Methods: Lecture 1

Meredith Fowlie
UC Berkeley

September 2021

Recall our structural estimation overview:

Five (not-so-)easy steps!!

- Write down a sensible theoretical model (e.g. the Cobb Douglas production function relating inputs and outputs).
- Be explicit about what we can and cannot observe.
- Think about the processes that give rise to "structural errors". Impose assumptions on the statistical properties of these errors.
- Derive statistical objects implied by the model (such as a log likelihood function or conditional moments).
- Bring these to the data and identify the parameters that best match or rationalize the data we observe (conditional on the assumed structure).

Discrete Choice Modeling: Basic Set-up

An agent (household/firm/consumer) makes a choice from among J possible actions or options.

Discrete choice model applicable to many choice contexts:

- Consumer choice of durables (cars, appliances, etc)
- Transportation mode choices
- Occupation choice
- Electricity/gas rate plans
- Neighborhood choice
- Firm's choice of production/capital equipment
- Firm entry/exit
- etc.

Discrete choice modeling with simulation

Simulation-based methods have transformed discrete choice modeling:

- Estimate choice probabilities
- Estimate choice models
- Simulate choices in counterfactual choice contexts.

Simulation-based methods allow researchers to write down the choice model that represents behavior without worrying about closed-form tractability.

- Approximate integrals using simulation (versus solving for closed-form analytic soultion).

Outline for today

- Discrete choice Random Utility Maximization (RUM)
- Introducing the logit model
- Estimation basics (SLL)
- Why OLS intuition can do you wrong in Logit-land
- Limitations of Logit
- Welfare measures and logit
- Mixed Logit

Discrete choice modeling

What kinds of choice situations are we thinking about?

- The dependent variable, Y, takes on non-negative, un-ordered integer values between zero and J.
- Typically motivated as an economic choice (i.e. some kind of constrained optimization) from a set of discrete options or alternatives.

Discrete choice modeling

Whatever the application/choice context, we will limit our attention to choice sets with the following three characteristics:

- Mutually exclusive states: $j = 1...J.$: The agent can choose only one of these states.
- Exhaustive: The values $j = 1..J$ should capture all possible realizations of the dependent variable (all choices represented).
- Finite.

Focus is on conditional probabilities

- Goal is to estimate/understand the conditional probability that agent n chooses option j conditional on observed choice characteristics X.
- Inevitably, there are factors we don't observe which also determine the agent's choice.
- We will spend much of this thinking about 'nuisance' parameters.

The Random Utility Model (RUM)

- In the economics literature, discrete choice models are typically derived from economic models of utlity maximization or cost minimization.
- The most standard framework: random utility maximization.
- Consider a utility maximizing individual choosing among J alternatives (e.g. alternative vehicle models, transporation modes, vacation destinations, careers, etc).
- Let U_{jn}^* be the utility that individual n derives from choosing alternative j.
- The individual will choose i if $U_{ni}^* > U_{nj}^*$ for all $j \neq i$.

Random Utility Model

- The individual-specific, choice-specific utility U_{ni}^* is a "latent" value in the sense that we cannot observe it directly.
- We can observe some covariates X that we believe play a role in determining the choice.
- But there are also factors implicated in the data generating process that we cannot observe.

We decompose latent utility into two parts:

$$U_{ni}^* = \underbrace{U(X_{ni}; \beta)}_{\text{deterministic component}} + \underbrace{\epsilon_{ni}}_{\text{unobserved component}} \tag{1}$$

Random Utility Model

$$U_{ni}^{*} = \underbrace{U(X_{ni};\beta)}_{\text{deterministic component}} + \underbrace{\varepsilon_{ni}}_{\text{unobserved component}} \qquad (2)$$

To approx nuisance params that affect peoples decisions (feels right) that we can't measure and are dependent on the person.

- The vector of residuals ε_n includes all of the J choice specific disturbances associated with individual n.
- It is convenient (although not necessarily accurate!) to assume these disturbances are independent and identically distributed both across agents and across choice situations.

How do we get from here to a formulation of conditional choice probabilities?

Conditional choice probabilities

Step 1: Formulate the conditional probability that agent n chooses alternative i:

$$\begin{aligned}\Pr(Y_n &= i|X_n = x) = \Pr(U^*_{ni} > U^*_{nj})\ \forall i \neq j \\ &= \Pr(U(X_{ni}) + \varepsilon_{ni} > U(X_{nj}) + \varepsilon_{nj}) \forall i \neq j \\ &= \Pr(\varepsilon_{nj} - \varepsilon_{ni} < U(X_{ni}) - U(X_{nj})) \forall i \neq j\end{aligned}$$

- This is the cumulative probability that the differences in the disturbances are less than the differences in the observed component.
- It will be useful to express this choice probabiity as something we can integrate over the assumed distribution of disturbances.

More machinery...

Individual n will choose option i over j if

$$I(\varepsilon_{nj} - \varepsilon_{ni} < U(X_{ni}; \beta) - U(X_{nj}; \beta)) = 1,$$

where I is an indicator function that equals 1 if the statement in parentheses is true, zero otherwise.
This formulation is convenient because it allows us to express the probability as an integral:

$$\Pr(Y_n = i | X_n = x) = \int_{\varepsilon_n} I(\varepsilon_{nj} - \varepsilon_{ni} < U(X_{ni}; \beta) - U(X_{nj}; \beta)) f(\varepsilon_n) d\varepsilon_n.$$

If we know the distribution of the disturbances $f(\varepsilon_n)$, the integral over the density of the unobserved portion of utility can be computed/estimated.

Let's work through an example

Refrigerators are one of the most energy intensive home appliances!

Consumer choice of refrigerator

There is much we can observe about this choice (= fun times with Consumer Reports!):

- price
- freezer location
- brand
- energy efficiency
- capacity

These are the X variables that determine the deterministic utility component $U(X_{in})$ each consumer derives from a given refrigerator choice.

Suppose, based on what we observe (and the structure we impose on $U(X_{nj}; \beta)$), we estimate that $U(X_{ni}; \beta) = 5$ and $U(X_{nj}; \beta) = 3$.

Modeling discrete (refrigerator) choices

- If our utility model is properly specified, and if agents make utility maximizing choices, and we can perfectly observe all determinants of this appliance choice, individual n will choose appliance model i.
- If there are components of this utility maximization that we cannot observe (likely), the best we can do is estimate the probability that individual n will choose option i conditional on X.

$$\Pr(Y_{nj} = i|X_n) = Pr(\varepsilon_{nj} - \varepsilon_{ni} < 5 - 3) \tag{3}$$

Punchline: People make choices in accordance with $U(X_{in}; \beta)$ unless the unobserved components reverse this reference ordering.

Structural assumptions in this discrete choice model?

To make this model empirically tractable, we need:

1. A plausible economic model of the underlying choice/optimization problem.
2. Assumptions about the statistical properties of the error distribution.

If we want an elegant *closed form* expression for the choice probabilities, we need to find an economic model and a disturbance distribution that play nicely together....

Recall our structural estimation overview:

Five (not-so-)easy steps!!

- Write down a sensible theoretical model (e.g. specify a utility maximization framework).
- Be explicit about what we can and cannot observe.
- Think about the processes that give rise to "structural errors". Impose assumptions on the statistical properties of these errors.
- Derive statistical objects from the model (in this case, a log likelihood function).
- Bring this LL to the data and identify the parameters that best match or rationalize the choices we observe (conditional on the assumed structure).

Outline for today

- Discrete choice and Random Utility Maximization
- **Introducing the logit model**
- Estimation basics (SLL)
- Limitations of Logit
- Why OLS intuition can do you wrong in Logit-land
- Welfare measures and logit
- Mixed Logit

Meet the Logit family

The logit family of models is recognized as an essential foundation for empirical discrete choice models.

- Conditional logit model(CL): Choice probabilities are a function of the attributes of the choice alternatives.
- Multinomial logit models (MNL): Choice probabiities are a function of the characteristics of the agent.

The CL model will provide a natural point of departure because it is most consistent with economic models of utility maximization/cost minimization.

Conditional logit model

- The logit model is a tractable model with elegant, closed form choice probabilities:

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

- The catch? It imposes strong assumptions that are not appropriate for every empirical setting.
- A good jumping off point as it provides the foundation/substrate for more general discrete choice models.

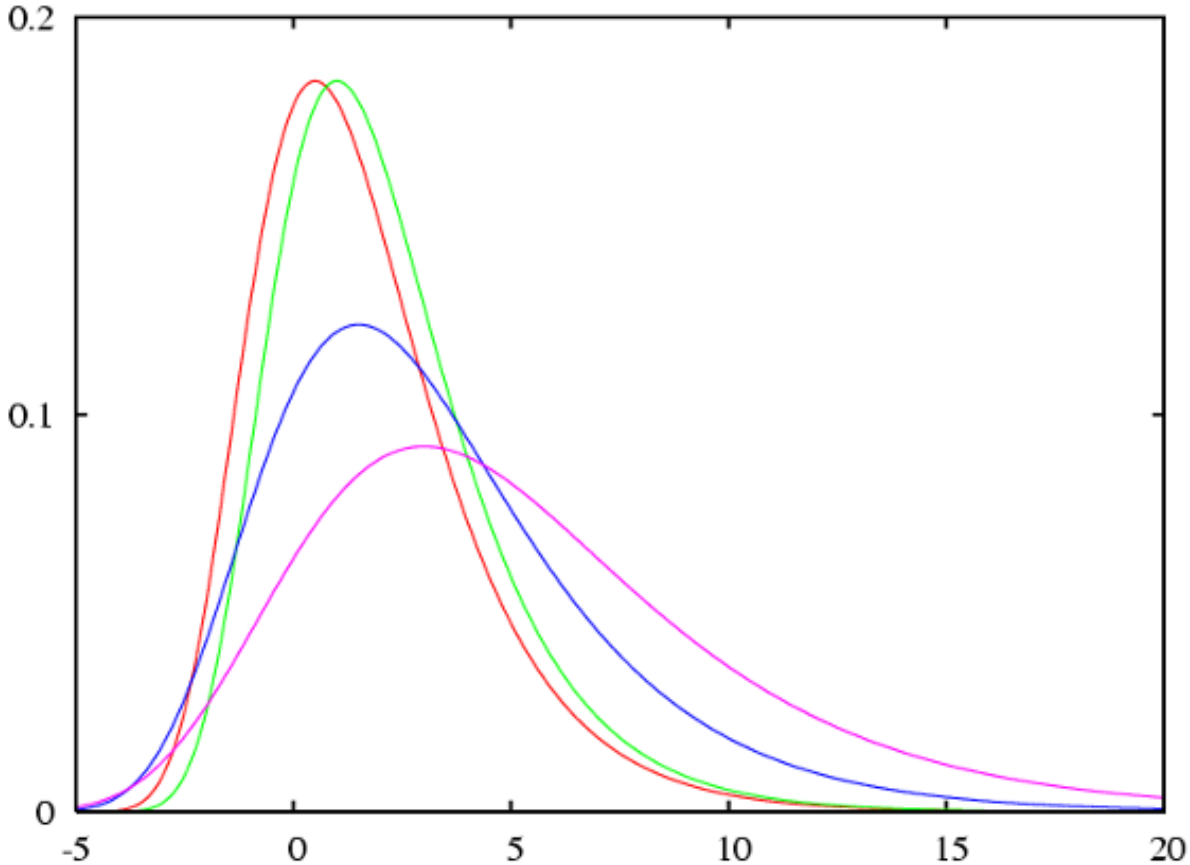
Conditional logit model

Basic set-up:

$$\mathrm{U}^*_{nj} = U(X_{nj}; \beta) + \varepsilon_{nj}$$
$$\varepsilon_{nj} \sim iidEV1$$

- Covariates X_{nj} vary across choices.
- Choices can also vary by indivudual.
- Unobserved component assumed to be $\sim$ iid distributed according to the extreme value distribution (similar to normal.. but fatter tails and asymmetric).
- Why the extreme value distribution? It yields analytically integrable conditional choice probabilities!

Meet the EV1 density



EV1 density defined by a location μ and scale parameter σ.
Distribution of the difference of two EV1 random variables has a logistic distribution.

Meet the EV1 density

- There are two parameters that define this distribution: the location parameter μ and the scale parameter σ.
- If we set $\mu = 0$ and $\sigma = 1$ we obtain the standard EV1 with a density: $f(x) = e^{-x}e^{-e^{-x}}$.
- The cumulative distribution is: $F(x) = e^{-e^{-x}}$.
- The difference between two *iid* EV1 distributed random variables has a logistic distribution. Let ε^* denote the difference between two EV1 distributed disturbances: $f(\varepsilon^*) = \frac{e^{-(x-\mu)/\sigma}}{\sigma(1+e^{-(x-\mu)/\sigma})}$
- The cumulative distribution of this difference: $F(\varepsilon^*) = \frac{1}{1+e^{-(x-\mu)/\sigma}}$.

Let's work through an application

- The standard way to motivate the conditional logit choice probabilities builds directly on a model of random utility maximization.
- CL model can also be motivated with a model of cost *minimization*.
- For additive, iid extreme value (Type I) errors, the assumption of cost minimization does not yield the standard CL choice probabilities due to the asymmetry of the assumed distribution.
- But if instead we assume that the extreme value term is *subtracted* from (versus added to) the deterministic component, we can derive an equally convenient expression.

CL meets cost minimization

- Cost minimizing firms (indexed by n) choose from among J production technology alternatives.
- Assume choice set meets the criteria we specified (mutually exclusive, exhaustive, finite).
- We observe covariates X_{nj} that presumably play a role in determining the costs as perceived by the firm:

$$C_{ni} = \beta_{0i} + \beta_1 LAC_{ni} + \beta_2 E_{ni} + \sum_{k=3}^{K} \beta_k X_{kni},$$

- LAC_{ni} is the levelized annual investment cost of the technology
- E_{ni} and X_{kni} are the annual expenditures on energy and k non-energy inputs.
- β_{0i} are technology specific constants.

CL meets cost minimization

The latent cost that the firm associates with each alternative is modeled as:

$$C_{ni}^{*} = \beta' X_{ni} - \varepsilon_{ni}$$

- Assume these ε_{ni} are independently and identically distributed type I extreme value with $\mu = 0$ and scale parameter $\sigma = 1$ (more on this scale parameter normalization below).
- Assume that the firm chooses the compliance option that minimizes anticipated compliance costs.

Derivation of conditional choice probabilities

Let P_{ni} be the probability that unit n chooses alternative i :

$$\begin{aligned} P_{ni} &= \text{Prob}\left(\beta' X_{ni} - \varepsilon_{ni} < \beta' X_{nj} - \varepsilon_{nj} \quad \forall\, j \neq i\right) \\ &= \text{Prob}\left(\varepsilon_{nj} < \beta' X_{nj} - \beta' X_{ni} + \varepsilon_{ni} \quad \forall\, j \neq i\right) \end{aligned}$$

The expression for the *conditional* choice probability :

$$\begin{aligned} P_{ni}|\varepsilon_n &= \prod_{j \neq i} F(\beta' X_{nj} - \beta' X_{ni} + \varepsilon_{ni}) \\ &= \prod_{j \neq i} \exp(-\exp(-(\beta' X_{nj} - \beta' X_{ni} + \varepsilon_{ni}))) \end{aligned}$$

Recall the EV1 cumulative distribution is: $F(x) = e^{-e^{-x}}$.

Unconditional choice probabilities

Unconditional choice probabilities are obtained by integrating over the distribution of ε_n :

$$P_{ni} = \int_{\varepsilon=-\infty}^{\infty} \prod_{j \neq i} \exp(-\exp(-(\beta' X_{nj} - \beta' X_{ni} + \varepsilon_{ni}))) f(\varepsilon_n)\ d\varepsilon_n$$

Make a bunch of clever substitutions and re-arrangements (see notes) to get here:

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

An alternative way of expressing this choice probability:

$$P_{ni} = \frac{\frac{1}{\exp(\beta' X_{ni})}}{\sum_j (\frac{1}{\exp(\beta' X_{nj})})} = \frac{\exp(-\beta' X_{ni})}{\sum_j \exp(-\beta' X_{nj})}$$

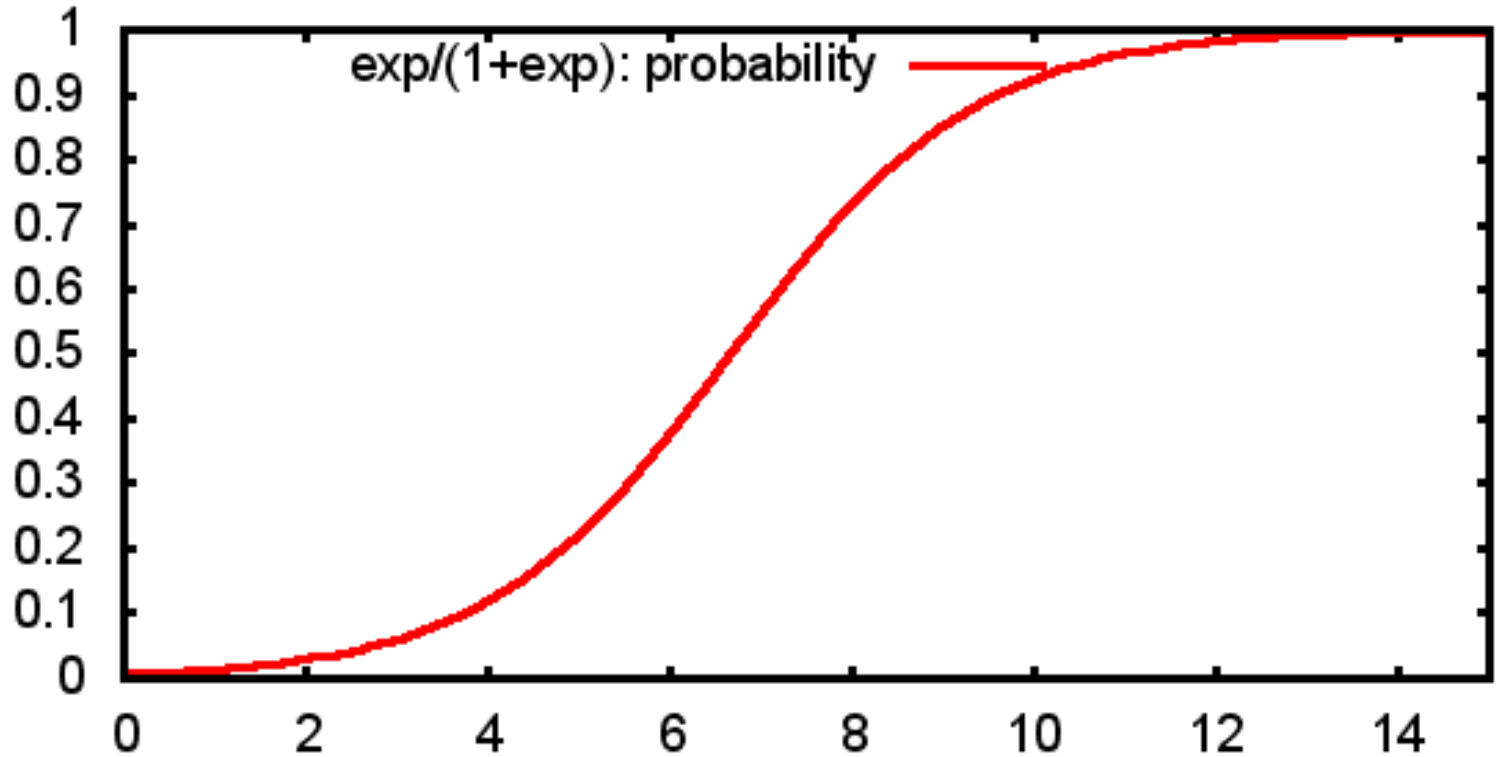
Lots to like about this conditional choice probability!

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

- Choice probabilities lie between 0 and 1 by design.
  - For very high costs, -$\beta' X_{ni}$ becomes a large negative number, and the numerator approaches zero.
  - For very low costs, -$\beta' X_{ni}$ approaches zero and numerator approaches one.
- Choice probabilities sum to one by design.

CL choice probabilities imply this:

The CL choice probability formula implies that the relationship between choice probabilities P_{ni} and an observable choice attribute X_{ni} follows a sigmoid shape.



This is intuitive.

Outline for today

- Discrete choice and Random Utility Maximization
- Introducing the logit model
- **Estimation basics (SLL)**
- Why OLS intuition can do you wrong in Logit-land
- Limitations of Logit
- Welfare measures and logit
- Mixed Logit

Estimation

To introduce the basics of estimation, we invoke several convenient assumptions (some of which we will release later on):

- All covariates are exogenous
- Our sample is either a cross section or a panel with zero serial correlation.
- Each decision maker's choice is independent of the choices of other decision makers.
- All alternatives are represented in the finite choice set.
- Disturbances are distributed *iid* extreme value.

These simplifying assumptions notwithstanding, estimating the parameters of a non-linear function is somewhat complicated.

Maximum Likelihood Estimation

Because the conditional choice probabilities have a non-linear closed-form, estimate the parameters of the model using maximum likelihood:

- Specify a likelihood function which defines the probability of choice outcomes as a function of the parameters we want to estimate.
- Find the parameter values that maximize the likelihood of observing the choices we observe.
- How to build our likelihood function?

Building our likelihood function

Let $y_{ni} = 1$ if agent n chooses i, and zero otherwise.

The probability of observing agent n making the choice y_{ni} that is actually observed:

$$\prod_i \left(\frac{\exp(\beta' X_{ni})}{\sum_j \exp(\beta' X_{nj})} \right)^{y_{ni}},$$

- P_{ni} raised to the power of $0 = 1$ for all alternatives not chosen, so this is simply the probability of selecting the chosen alternative.
- Given the assumed independence of decisions across decision makers, the probability that each agent in the data chooses the alternative she chose:

$$L(\beta) = \prod_{n=1}^{N} \prod_{j=i} \left(\frac{\exp(\beta' X_{ni})}{\sum_j \exp(\beta' X_{nj})} \right)^{y_{ni}}$$

Building our log likelihood function

$$L(\beta) = \prod_{n=1}^{N} \prod_{i} \left(\frac{\exp(\beta' X_{ni})}{\sum_j \exp(\beta' X_{nj})} \right)^{y_{nj}}$$

The log likelihood function is thus:

$$\begin{aligned} LL(\beta) &= \sum_n \sum_i y_{ni} \ln \left(\frac{\exp(\beta' X_{ni})}{\sum_j \exp(\beta' X_{nj})} \right) \\ &= \sum_n \sum_i y_{ni} (\beta' X_{ni}) - \sum_n \sum_i y_{ni} \ln \left(\sum_j \exp(\beta' X_{nj}) \right) \end{aligned}$$

Maximum Likelihood Estimation

$$LL(\beta) = \sum_n \sum_i y_{ni}(\beta' X_{ni}) - \sum_n \sum_i y_{ni} \ln \left(\sum_j \exp(\beta' X_{nj}) \right)$$

- ML estimates of β are those that maximize LL(β).
- Conditional on the assumptions we have made, this is a well behaved, globally concave function.
- At the maximum of the likelihood function, the derivative with respect to each of the parameter values is zero: $dLL(\beta)/d\beta = 0$.

Why will the value of our LL function be negative??

Goodness of fit?

Percent correctly predicted

- Define a "prediction success indicator" for each agent (e.g. $s_n = 1$ if $y_{nj} = i$.
- Define percent correctly predicted: $\frac{\sum_n s_n}{N}$.
- Problematic in that it is premised on the idea that the best prediction for each person is the choice with the highest probability.

Likelihood ratio test

- Calculate the value of the LL function evaluated at $\hat{\beta_{MLE}}$.
- Calculate the value of the LL function evaluated at $\beta = 0$ (or a model with choice dummies only).
- LR tells you the percent improvement at convergence over no-information (or limited information) model:

$$LR = \frac{LL(0) - LL(\beta)}{LL(0)}.$$

Outline for today

- Discrete choice and Random Utility Maximization
- Introducing the Logit model
- Estimation basics (SLL)
- **Why OLS intuition can do you wrong in Logit-land**
- Limitations of Logit
- ~~Welfare measures and logit~~
- Mixed Logit

Logit ≠ OLS!!

1. Everything is relative
2. Coefficients are hard to intepret out of the box.
3. These are not your OLS' interaction effects.
4. Coefficient comparisons across groups are tricky (impossible?).

↳ comparing utils to utils

It's all relative!

Only relative differences in utility matter:

$$\begin{aligned} \mathrm{U}_{1n}^{*} &= U(X_{1n}) + \varepsilon_{1n} \\ \mathrm{U}_{2n}^{*} &= U(X_{2n}) + \varepsilon_{2n} \end{aligned}$$

Add a constant c to both latent utility values, choice behavior unchanged.

It's all relative!

Only relative differences in utility matter:

$$\mathrm{U}_{1n}^{*} = U(X_{1n}) + \varepsilon_{1n}$$
$$\mathrm{U}_{2n}^{*} = U(X_{2n}) + \varepsilon_{2n}$$

Add a constant c to both latent utility values, choice behavior unchanged.

One intuitive implication: you cannot estimate coefficients on attributes/characteristics that do not vary across choices (e.g. gender).

It's all relative!

$$\begin{aligned} U^*_{1n} &= \delta_1 + \beta'(X_{1n}) + \varepsilon_{1n} \\ U^*_{2n} &= \delta_2 + \beta'(X_{2n}) + \varepsilon_{2n} \end{aligned}$$

The δ are not identified.. you need to normalize:

$$\begin{aligned} U^*_{1n} &= \beta'(X_{1n}) + \varepsilon_{1n} \\ U^*_{2n} &= \delta_2 + \beta'(X_{2n}) + \varepsilon_{2n} \end{aligned}$$

What does this δ_2 capture?

It's all relative!

$$\begin{aligned} U^*_{1n} &= \delta_1 + \beta'(X_{1n}) + \varepsilon_{1n} \\ U^*_{2n} &= \delta_2 + \beta'(X_{2n}) + \varepsilon_{2n} \end{aligned}$$

The δ are not identified.. you need to normalize:

$$\begin{aligned} U^*_{1n} &= \beta'(X_{1n}) + \varepsilon_{1n} \\ U^*_{2n} &= \delta_2 + \beta'(X_{2n}) + \varepsilon_{2n} \end{aligned}$$

What does this δ_2 capture?

It tells us something about the average effect of unobserved attributes of choice 2 relative to the unobserved attributes of the omitted choice.

Discrete choice models are only identified up to scale

Only relative differences in utility matter:

$$\begin{aligned} \mathsf{U}_{1n}^{*} &= U(X_{1n}) + \varepsilon_{1n} \\ \mathsf{U}_{2n}^{*} &= U(X_{2n}) + \varepsilon_{2n} \end{aligned}$$

Multiply through by a scale factor s, choice behavior unchanged.

So what?

Discrete choice models are only identified up to scale

Only relative differences in utility matter:

$$\begin{aligned} \mathrm{U}_{1n}^{*} &= U(X_{1n}) + \varepsilon_{1n} \\ \mathrm{U}_{2n}^{*} &= U(X_{2n}) + \varepsilon_{2n} \end{aligned}$$

Multiply through by a scale factor s, choice behavior unchanged.

So what?

- If the choices we observe are unaffected by the scale of the covariates in the model, the scale of the β coefficients is not identified!
- This has important implications for how we specify discrete choice models *and* how we interpret our coefficient estimates.

Identification requires normalization!

- Discrete choice model must be normalized to take account of the fact that the level and scale of the latent LHS value are not identified.
- If the error terms are assumed to be *iid*, then the scale normalization is relatively straightforward: Normalize the error variance to some number.
- When the observed portion of utility is linear in parameters, the normalization provides a way of interpreting coefficients.
- Let's work through a standard normalization.

A standard normalization

- For the EV1 distribution, the variance of the error:

$$var(\varepsilon_{in}) = \sigma^2 \pi^2/6, \tag{4}$$

- Reformulate your choice model to divide through by σ :

$$\begin{aligned} U(X) &= \frac{\beta}{\sigma} X_{ni} + \frac{\varepsilon_{ni}}{\sigma} \\ &= \frac{\beta}{\sigma} X_{ni} + \varpi_{ni} \end{aligned}$$

Taste coef.s

  Now, the variance of this ϖ_{ni} disturbance is $\pi^2/6$

- Having pinned down the variance to equal $\pi^2/6$, the model is identified. BUT, the covariate coefficients should be interpreted as $\frac{\beta}{\sigma}$.

Normalization of the CL model

Return to our cost minimization example:

- Assume that the decision maker will choose option that minimizes costs:

$$\beta' X_{ni} - \varepsilon_{ni} \leqslant \beta' X_{nj} - \varepsilon_{nj} \ \forall j \neq i,$$

$$\varepsilon_{nj} \tilde{} EV(0, \frac{\pi^2}{6}\sigma^2)$$

- Assume that the residuals are distributed *iid* extreme value with $\mu = 0$.
- The variance of ε is $\frac{\pi^2}{6}\sigma^2$, where σ is the scale parameter of the EV1 distribution.

Normalization of the CL model

Normalize the variance: divide through by the scale parameter σ:

$$\frac{C_{nj}}{\sigma} = \frac{\beta'}{\sigma} X_{nj} + \frac{\varepsilon_{nj}}{\sigma} \equiv \frac{\beta'}{\sigma} X_{nj} + \varpi_{nj}$$

The normalized variance:

$$var(\varpi_{nj}) = var(\frac{\varepsilon_{nj}}{\sigma}) = \frac{1}{\sigma^2}\frac{\pi^2}{6}\sigma^2 = \frac{\pi^2}{6}$$

All β coefficients are now normalized by the scale parameter.

We are not in OLS land anymore!

Suppose we estimate the relationship between firm-level energy consumption Y and observable variables such as plant operating capacity x_1 and age x_2:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\frac{\partial E[Y]|x_1, x_2]}{\partial x_1} = \beta_1$$

β_1 has a very intuitive interpretation.

Conditional logit coefficients?

Suppose we analyze how projected annual savings less levelized annual investment costs determine a firms' decision to adopt a new technology:

$$\begin{aligned} V_{nj} &= \beta_1 SAV_{nj} - \beta_2 COST_{nj} - \varepsilon_{nj} \\ & \\ & \varepsilon_{nj} \tilde{} EV(0, \frac{\pi^2}{6}) \end{aligned}$$

How to interpret our estimated (e.g. MLE) β coefficients??

Coefficients hard to interpret out of the box

The logit choice probabilities in this binary choice context are:

$$\Pr(Y = 1|X) = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}$$

- β measures the partial effect of an incremental change in the corresponding covariate x on the latent dependent variable scaled by the unknown/unidentified σ parameter.
- β tells us how important X is *relative to* unobserved factors.
- If the estimated β is large (small), this suggests that variation in costs explain much (little) of the choice variation relative to unobserved factors.

Because these β coefficients are so hard to interpret, it is standard to report more intuitive *implications* of these estimates.

Marginal effects (math)

Marginal effects capture how the choice probabilities are affected by an incremental change in a given covariate x.

$$E[Y|x_i, x_2] = \frac{\exp(\beta' X_{ni})}{\sum_j \exp(\beta' X_{nj})}$$

$$\frac{dP_{ni}}{dx_{1i}} = \frac{(\beta_1)\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{\sum_j \exp(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j})} - \frac{(\beta_1)\left(\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})\right)^2}{\left(\sum_j \exp(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j})\right)^2}$$

$$= (\beta_1)(P_{ni} - (P_{ni})^2)$$

$$= (\beta_1)P_{ni}(1 - P_{ni})$$

Marginal effects (intuition)

- If the latent outcome function is linear, the marginal effect of a change in x_1 on the probability that an outcome will occur is obtained by multiplying the coefficient β_1 by $P_{ni}(1 - P_{ni})$.
- This is maximized when $P_{ni} = 0.5$.
- Intuitively, the effect of an incremental change in an attribute value on the underlying choice probabilities is highest when the agent is on the fence.
- We can also evaluate how the choice probability P_{ni} is affected by an incremental change in an attribute of another alternative j:

$$\frac{dP_{ni}}{dx_{nj|}} = -P_{ni}P_{nj}$$

Elasticities

The elasticities implied by the CL parameter estimates provide another intuitive way to present estimation results.

The elasticity of P_{ni} with respect to X_{ni}:

$$\epsilon_{iX_{ni}} = \beta_1(1 - P_{ni})X_{ni}$$

The elasticity of P_{ni} with respect to X_{nj}:

$$\epsilon_{ix_{nj}} = -\beta_1 P_{nj} x_{nj}$$

Note that the effect of a change in attribute x_{nj} changes the probabilities for all other choices by the same percent.

Logit $\neq$ OLS!

1. It's all relative
2. Coefficients hard to interpret out of the box.
3. Interaction effects are not what you think they are.
4. Coefficient comparisons across groups are not what you think they are.

Interaction terms

In linear models, we often include interaction terms to infer how the effect of X_1 on Y depends on X_2.

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \\
\frac{\partial E[Y]}{\partial x_1} &= \beta_1 + \beta_3 x_2 \\
\frac{\partial E[Y]}{\partial x_2} &= \beta_2 + \beta_3 x_1 \\
\frac{\partial E[Y]^2}{\partial x_1 \partial x_2} &= \beta_3
\end{aligned}
$$

This intuition does not map perfectly into a discrete choice framework..Why?

Interaction terms

$$U^*_{nj} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{1j} x_{2j} + \varepsilon_{nj}$$
$$\varepsilon_{nj} \sim iidEV1$$

$$P_{ni} = \frac{\exp(\beta' X_{ni})}{\sum_j \exp(\beta' X_{nj})} \tag{5}$$

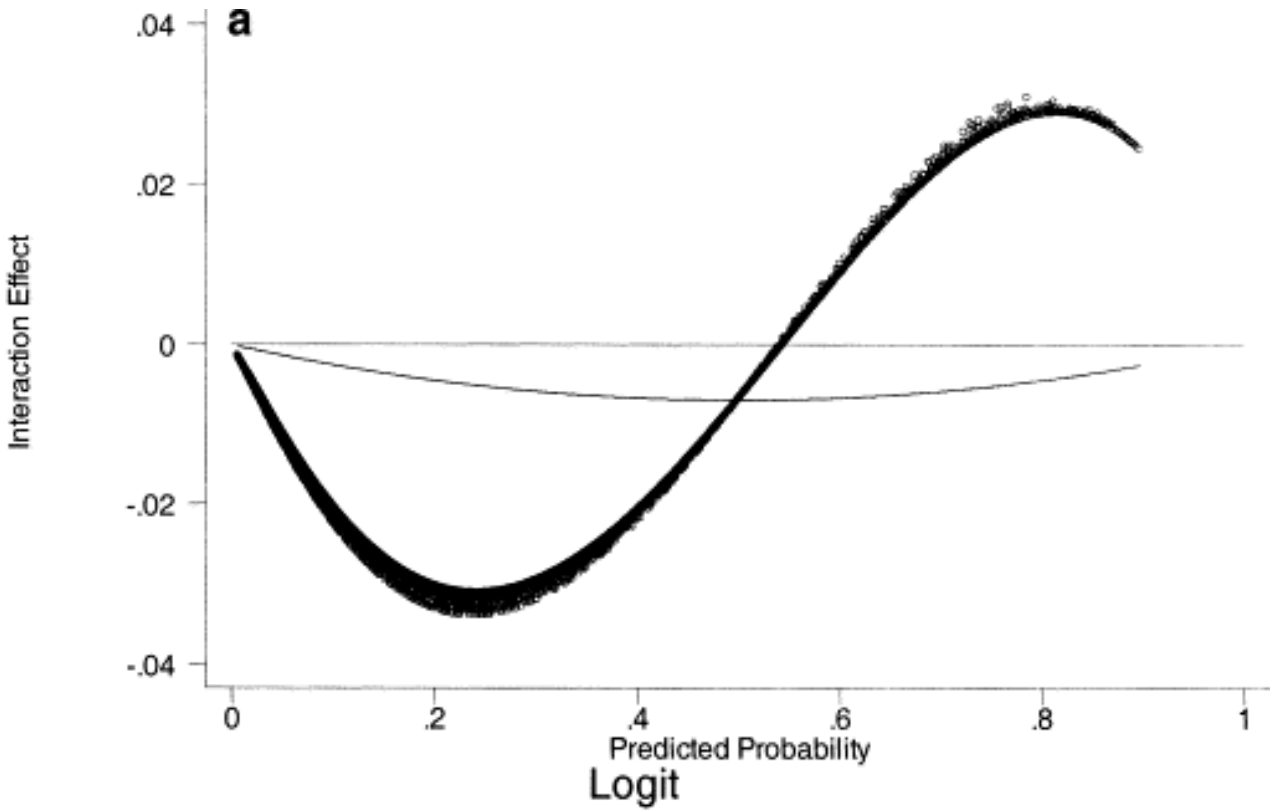
- Interaction effects on choice probabilities depend on the values of *all* covariates, so estimated interactions effect can vary significantly across units in the data.

Punchline: Interaction effects, often a variable of interest in applied econometrics, cannot be evaluated simply by looking at the sign, magnitude, or statistical significance of the coefficient on the interaction term when the model is nonlinear.

A simple example (from Ai and Norton, 2003)

- Consider a simple logit model that predicts HMO enrollment as a function of age, number of activities of daily living (a count of the number of basic physical activities a person has trouble performing), and the percent of the county population enrolled in a HMO
- The dependent variable is a dummy variable indicating whether the individual is enrolled in a HMO.
- An interaction term between age and the number of limiting activities is negative but not statistically significant.
- One might be tempted to conclude that the effect of age on the HMO enrollment choice does not vary when you incrementally increase the number of limited activities...

Scatter plot provides a more accurate picture



The underlying interaction effects, which vary across individuals with different choice attributes X, is positive for many observations. And significant for many.

Logit $\neq$ OLS!

1. It's all relative
2. Coefficients hard to interpret out of the box.
3. Interaction effects are not what you think they are.
4. Coefficient comparisons across groups are not what you think they are.

Comparisons across sub-groups

$$\begin{aligned} U(X) &= \frac{\beta}{\sigma} X_{ni} + \frac{\varepsilon_{ni}}{\sigma} \\ &= \frac{\beta}{\sigma} X_{ni} + \varpi_{ni} \end{aligned}$$

- Discrete choice model covariate coefficients are confounded with unobserved heterogeneity in the residual variance.
- Under homoskedasticity, this implication of the standard logit model normalization is fairly harmless.
- If the residual variance differs systematically across observations (heteroskedascity) we need to be more mindful.
- Differences in the degree of residual variation across groups can produce apparent differences in slope coefficients that are not indicative of true differences in marginal effects.

Comparisons across sub-groups

Suppose we want to know whether the binary decision to accept an efficiency retrofit varies across types of firms.

- Tests of this hypothesis can be formulated as a test of whether the coefficients on an interaction between the covariates in a latent cost model and a group dummy are statistically significant.
- Unfortunately, these test results will be difficult to interpret.
- If the residual variance differs across groups, the coefficient estimates to be compared are confounded with scale factor differences.

Work arounds??

Work arounds?

- Note that the scale parameter cancels when you construct group-specific ratios of coefficients.
- Good news: A ratio of two coefficients can be compared directly across groups.
- Bad news: If the supports of the estimated distributions of the parameters to be compared overlap zero, the variance of this ratio will not be well defined.
- One common solution involves assuming that the coefficient in the denominator is fixed (Layton and Brown, 2000).
- Sonnier and Train. (2005) offer some cautionary notes wrt this strategy.

Outline for today

- Discrete choice and Random Utility Maximization
- Introducing the logit model
- Estimation basics (SLL)
- Why OLS intuition can do you wrong in Logit-land
- **Limitations of Logit**
- Welfare measures and logit
- Mixed Logit

Limitations of the Conditional Logit

The *iid* error assumption is convenient .. but restrictive.

1. Preference heterogeneity
2. Repeated choices/panel data
3. Independence of irrelevant alternatives

These three limitations serve to motivate the mixed logit model.

Preference heterogeneity

- Different types of agents value choice attributes differently.
- We are often interested in modeling these heterogeneous preferences.
- Logit models can accommodate systematic taste variation.
- Logit models cannot accommodate random taste variation.

Systematic taste variation

Consider the latent cost function:

$$C_{ni} = \beta_{0i} + \beta_1 LAC_{ni} + \beta_2 ECOST_{ni} + \sum_{k=3}^{K} \beta_k X_{kni} + \varepsilon_{ni}$$

We might expect that different firms will weigh energy costs differently:

$$\beta_{2n} = \alpha + b(SIZE_n)$$

Making the substitution:

$$C_{ni} = \beta_{0i} + \beta_1 LAC_{ni} + \alpha ECOST_{ni} + b(SIZE_n) ECOST_{ni} + \sum_{k=3}^{K} \beta_k X_{kni} + \varepsilon_{ni}$$

No problem (although careful intepreting those interactions!!

Random taste variation?

Suppose the coefficient varies systematically and randomly across firms:

$$\beta_{2n} = \alpha + b SIZE_n + \nu_n$$

Making the substitution:

$$\begin{aligned} C_{ni} &= \beta_{0i} + \beta_1 LAC_{ni} + (\alpha + b SIZE_n + \nu_n) ECOST_{ni} + \sum_{k=3}^{K} \beta_k X_{kni} + \varepsilon_{ni} \\ &= \beta_{0i} + \beta_1 LAC_{ni} + \alpha ECOST_{ni} + b SIZE_n * ECOST_{ni} + \sum_{k=3}^{K} \beta_k X_{kni} + \mu_{ni} \\ \mu_{ni} &= \varepsilon_{ni} + \nu_n ECOST_{ni} \end{aligned}$$

- The error term now includes an interaction between ν_n and the choice-specific energy costs $ECOST_{ni}$.
- If energy costs are correlated across choices or individuals likely), the errors are no longer homoskedastic.

Panel data

- Suppose you are fortunate enough to have data on repeated choices made by the same decision maker.
- The CL model cannot accommodate correlation in the unobserved component of the latent utility/costs across choices made by the same agent.
- The CL model also cannot allow for state dependence/errors correlated over choices made by the same agent.

The Infamous IIA Property

- The standard logit implies substitution patterns that can be extremely unrealistic in some situations.
- To understand this property (and where it comes from) let's consider two features of the CL choice probabilities.
- First, note the ratio of two choice probabilities:

$$\frac{P_{ni}}{P_{nk}} = \frac{\frac{\exp(\beta' X_{ni})}{\sum_j \exp(\beta' X_{nj})}}{\frac{\exp(\beta' X_{nk})}{\sum_j \exp(\beta' X_{nj})}} = \frac{\exp(\beta' X_{ni})}{\exp(\beta' X_{nk})}. \tag{6}$$

- The relative odds of choosing i over k does not depend on the attributes of other available choices.
- So what? Why should the utility of one choice versus another depend on other choices?

Understanding IIA

Recall the cross-elasticities which capture how the probability of choosing alternative i is affected by a percentage change in an attribute of another alternative x_j.

$$\epsilon_{ix_j} = \frac{\frac{\partial P_{ni}}{P_{ni}}}{\frac{\partial x_{nj}}{x_{nj}}} = \frac{\partial P_{ni}}{\partial x_{nj}} \frac{x_{nj}}{P_{ni}} \tag{7}$$

$$= -(\beta_1)(P_{nj})\, x_{1j} \tag{8}$$

Note that this cross-elasticity of the choice probability P_{ni} with respect to the attribute x_{1j} is a function of the choice probability P_{nj} and the attribute x_{1j}.

This is the same percentage change for all choices not equal to j.

An illustrative example

Transport yourself back to 2017. Suppose you are trying to predict how the arrival of the Tesla-for-the-common-woman (model 3) will affect demand for other cars.

Consider a stylized example where the vehicle market is comprised only of the Toyota Tundra(70%) and Chevy Volt(30%)

- The logit model imposes that the ratio of Volt/Tundra market shares is unchanged (i.e. Tesla draws proportionally from these two very different vehicle market shares)
- This implies very unrealistic substitution patterns (and misleading projections wrt impacts on fuel consumption, etc.)

What's going on with IIA?

- With CL estimation, we are not uncovering substitution patterns in the data, the structure of the model is imposing them!
- The *iid* error assumes that unobserved components are identically distributed.
- The key to generating more realistic substitution patterns lies in specifying more realistic correlation in the unobserved components.

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.