The 14th Behavior Modeling Summer School

September 25-27, 2015

Introduction to Discrete Choice Models

Giancarlos Troncoso Parady
Assistant Professor
Urban Transportation Research Unit Department of Urban Engineering

THE UNIVERSITY OF TOKYO

Choice theory framework

Outcome of a sequential decision-making process:

1. Definition of the choice problem	→ *Choose a commuting mode*
2. Generation of alternatives	→ *Available modes: Car, transit, bike, walk*
3. Evaluation of attributes of the alternatives	→ *Weigh each alternative's attributes*
4. Choice	→ *Choose a mode*
5. Implementation	→ *Commute to work using the chosen mode*

This process defines the following elements:

1. Decision maker
2. Alternatives
3. Attributes of alternatives
4. Decision rule

Decision maker

- Individual, household, organization (i.e. firms, government agency)

Alternatives

$$Choice\ set \in \boldsymbol{Universal\ set}$$

Feasible alternatives known during the decision process

Defined by the environment of the decision maker

Alternative attributes

- A vector of characteristics that measure the attractiveness of an alternative *(e.g. Cost, comfort, travel time, etc)*

Decision rule

- Mechanism that defines the decision making process *(Dominance, satisfaction, lexicographic rules, **Utility**)*

An utility-maximization decision rule

- Attractiveness is reduced to a **single scalar function**
- Based on the notion of **tradeoffs**, or compensatory offsets, when making a choice.

- Assumption of **rational behavior:**
  - Under identical circumstances, an individual will repeat the same choices every time.

- **Random utility** approach:
  - Why? Because of observational deficiencies by the analyst, mainly a result of:
    1. Unobserved attributes
    2. Unobserved taste variations (heterogeneity)
    3. Measurement errors and imperfect information
    4. Proxy variables

An utility-maximization decision rule

- We can specify a random utility function as

$$U_{in} = V_{in} + \varepsilon_{in}$$

Observable (systematic) component — V_{in}

Unobservable (random) component — ε_{in}

So that

$$P(i|C_n) = \Pr(U_{in} > U_{jn}, \forall j \in C_n)$$

$$P(i|C_n) = \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \forall j \in C_n)$$

Only difference in utility matters!

$$= \Pr\left(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}, \forall j \in C_n\right) = \Pr(\varepsilon_n \leq V_n, \forall j \in C_n)$$

Where C_n is a feasible choice set for individual n

- To derive a specific model, we then need assumptions on

$$\varepsilon_{jn}, \forall j \in C_n$$

An utility-maximization decision rule

- Specifying the utility function components

$$U_{in} = V_{in} + \varepsilon_{in}$$

- Usually linear-in-parameters specification:

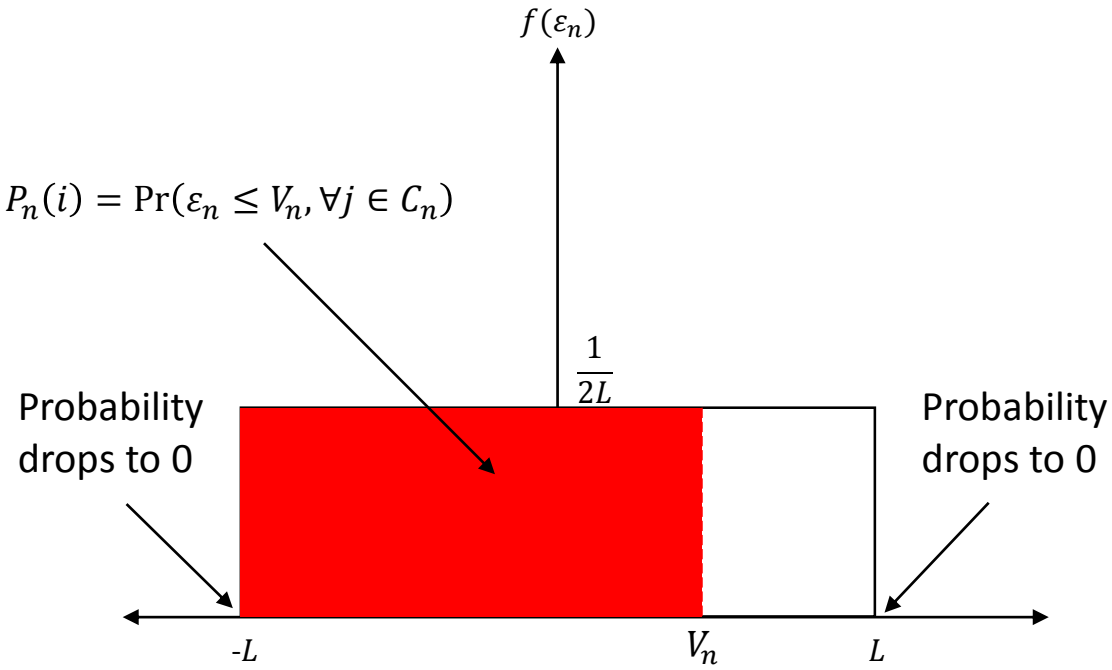
$$V_{in} = \beta_1 x_{in1} + \beta_2 x_{in2} + \cdots + \beta_K x_{inK}$$

where $x_{in} = f(z_{in}, S_n)$

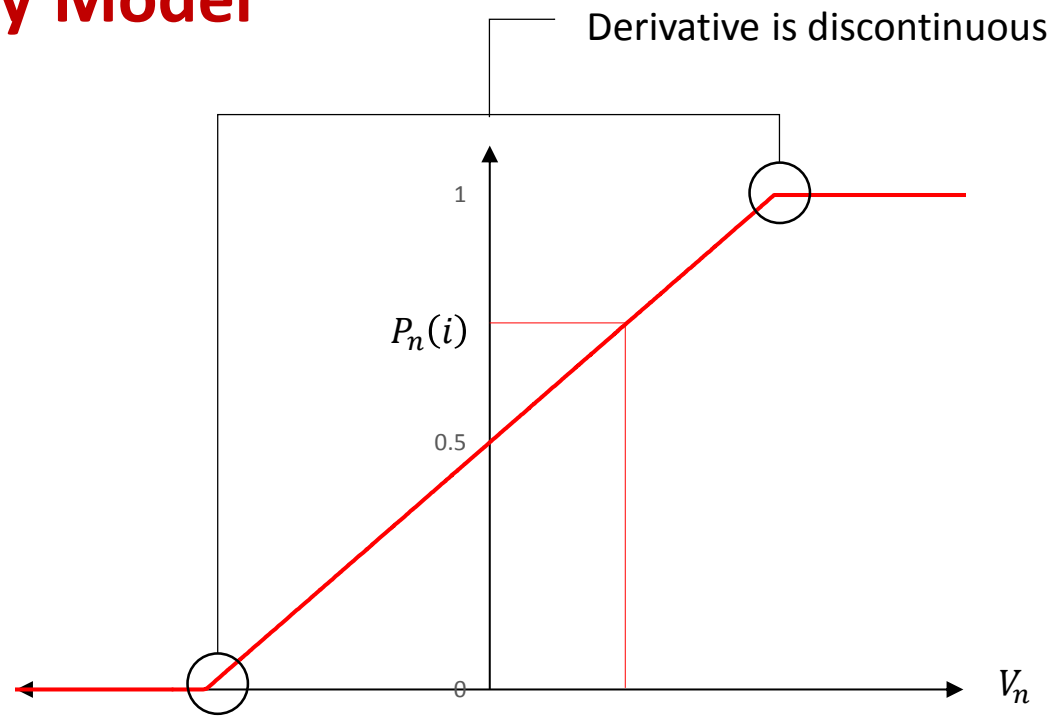
- Non-linearities can be introduced by allowing for any function f (polynomial, logarithmic, exponential, etc)

- Reflects the sources of randomness discussed earlier
- Different distributional assumptions result in different models:
  - *Normal distribution→Probit model*
  - *Gumbel distribution→Logit model*

Binary choice models: Linear Probability Model



Uniform distribution PDF of ε_n
(Our assumption about the error distribution)



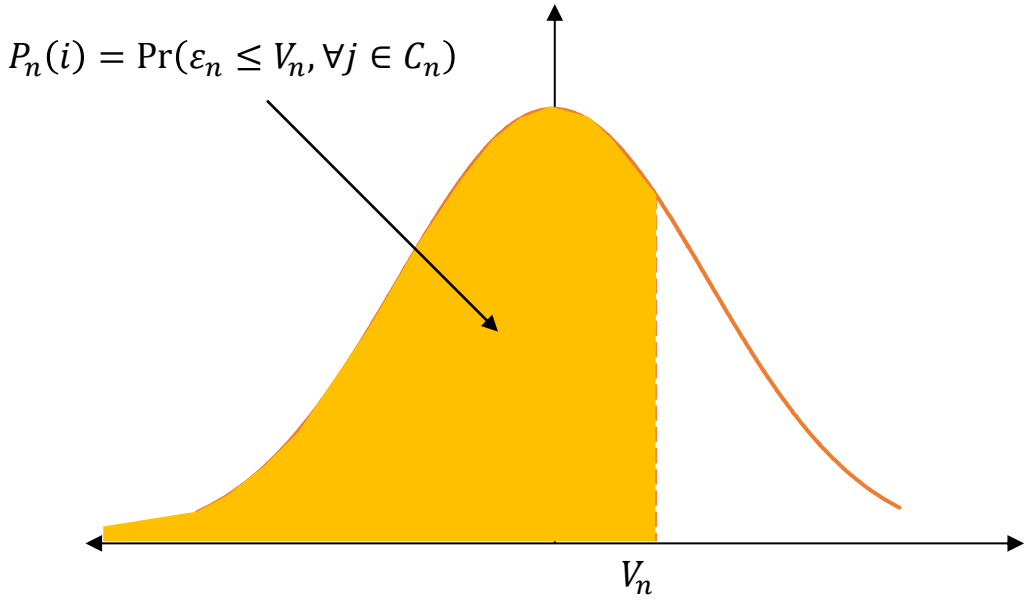
- The choice probability of i is given by the CDF of ε_n

$$P_n(i) = \begin{cases} 0 \\ \int_{-L}^{V_n} f(\varepsilon_n) d\varepsilon_n = \frac{V_n + L}{2L} \\ 1 \end{cases}$$

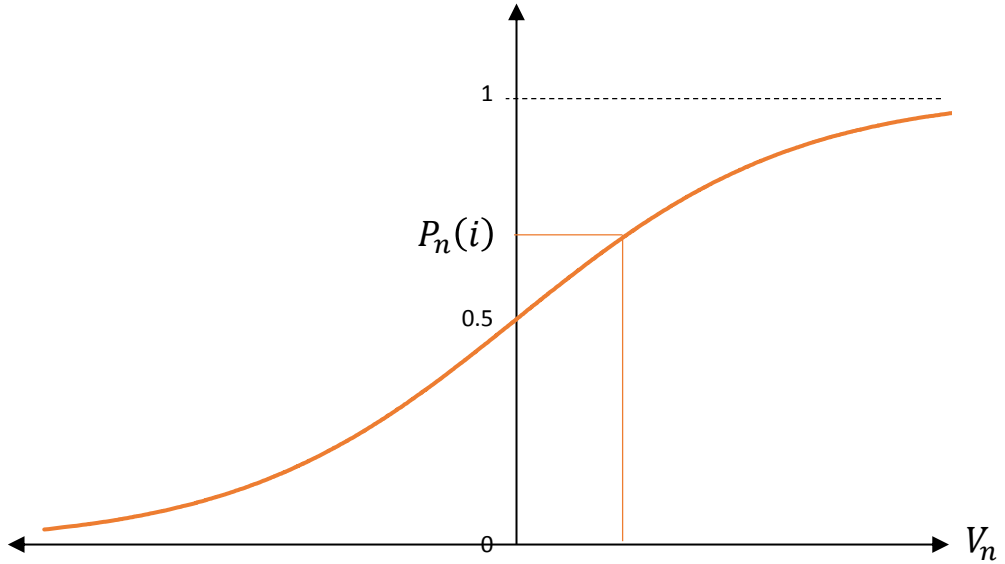
$$if\ V_n < -L$$
$$if\ -L \leq V_n \geq L$$
$$if V_n > L$$

Choices with predicted probability of 0 are still chosen.

Binary choice models: Probit Model



Normal distribution PDF of ε_n
(A better assumption about the error distribution)



- The choice probability of i is given by the CDF of ε_n

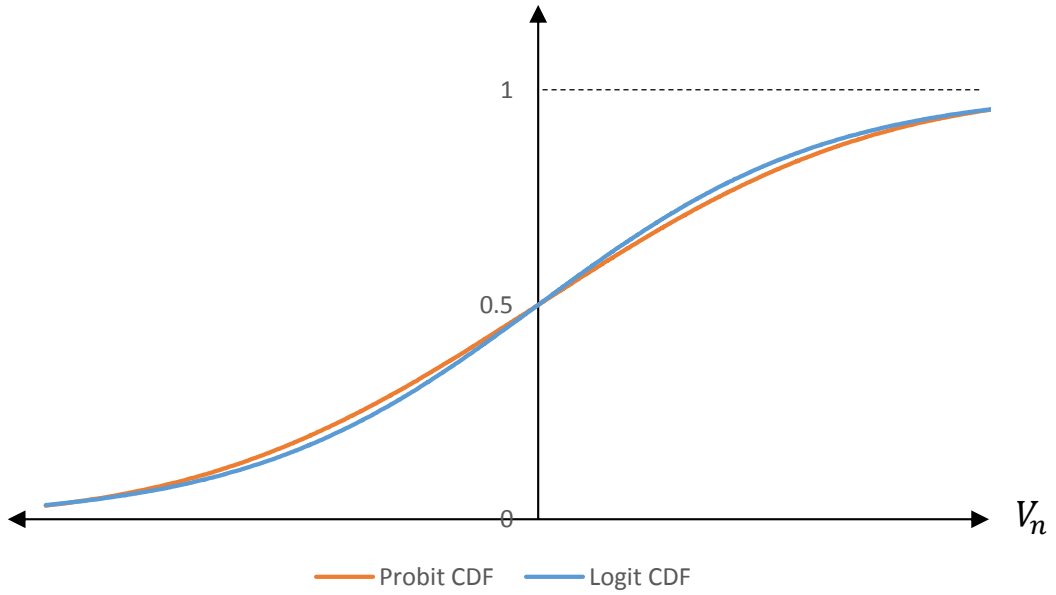
$$P_n(i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(V_n)/\sigma} exp\left[-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right] d\varepsilon = \Phi\left(\frac{(V_n)}{\sigma}\right)$$

Probabilities are never zero or one.

But the **probabilities cannot be expressed in a closed form** (numerical methods are required)

Binary choice models: Logit model

- A **probit-like model** that approximates a normal distribution.
- Probabilities **can be expressed in closed form**, so it is analytically convenient.
- ε_{in} and ε_{jn} are assumed to be **i.i.d. Gumbel distributed** (Type I extreme value distribution)
- So $\varepsilon_n = \varepsilon_{in}$-$\varepsilon_{jn}$ is **logistically distributed.**



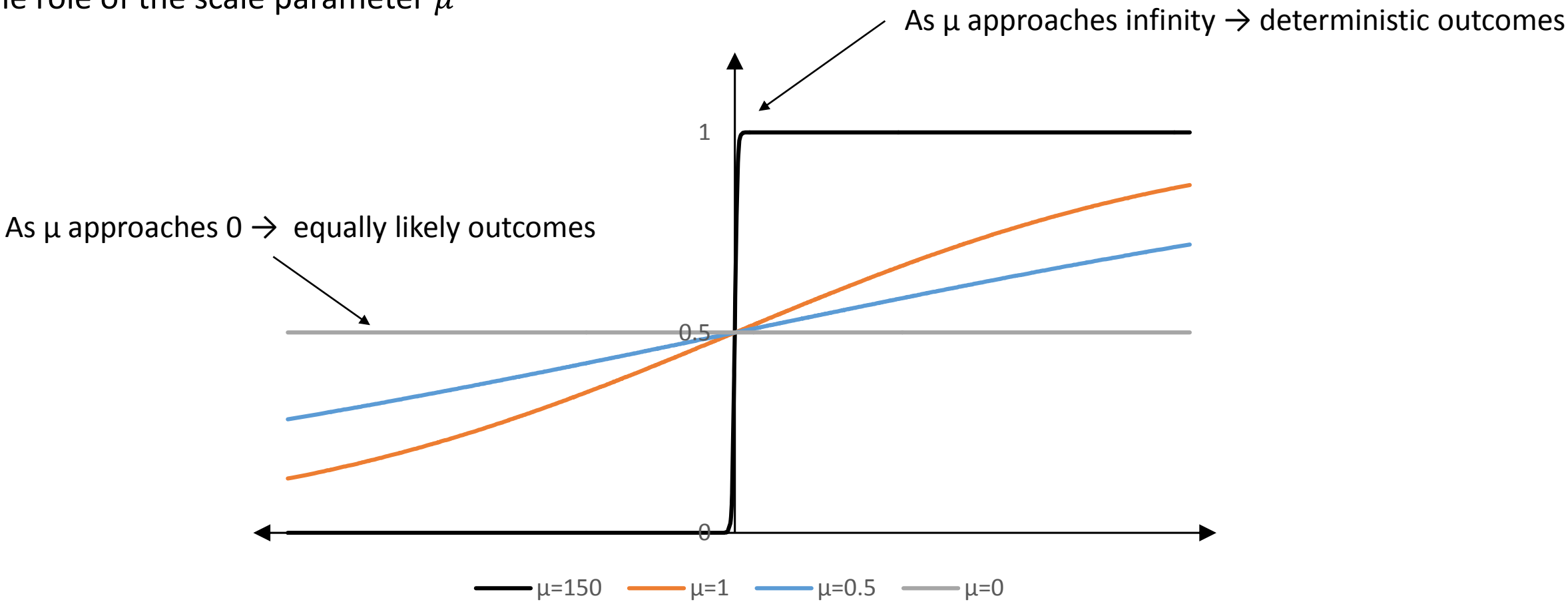
- The choice probability of i is given by the CDF of ε_n

$$P_n(i) = \frac{\exp(\mu V_{in})}{\exp(\mu V_{in}) + \exp(\mu V_{jn})} = \frac{1}{1 + \exp\left(-\mu(V_{in} - V_{jn})\right)}$$

where μ is a scale parameter

Binary choice models: Logit model

The role of the scale parameter μ



$$P_n(i) = \frac{\exp(\mu V_{in})}{\exp(\mu V_{in}) + \exp(\mu V_{jn})} = \frac{1}{1 + \exp\left(-\mu\left(V_{in} - V_{jn}\right)\right)}$$

Maximum likelihood estimation of parameters

The Maximum Likelihood principle states that, out of all the possible values of a parameter β, **the value that makes the likelihood of the observed data largest should be chosen.** (Wooldridge, 2004)

General form of the likelihood function:

$$L_n(\beta|x) = \prod_{n=1}^{N} f(x_n|\beta)$$

The likelihood is proportional the product of individual probabilities

Maximization of the Log-likelihood function

$$Max\ LL = \max_{\hat{\beta}_n} \sum_{n=1}^{N} logf(x_n|\beta)$$

Maximum likelihood estimation of parameters

In the general binary model case, the likelihood function can be defined as

$$L_n(\beta_1, \beta_2, \dots \beta_K) = \prod_{n=1}^{N} P_n(i)^{y_{in}} P_n(j)^{y_{jn}}$$

And the log-likelihood function we want to maximize can be defined as

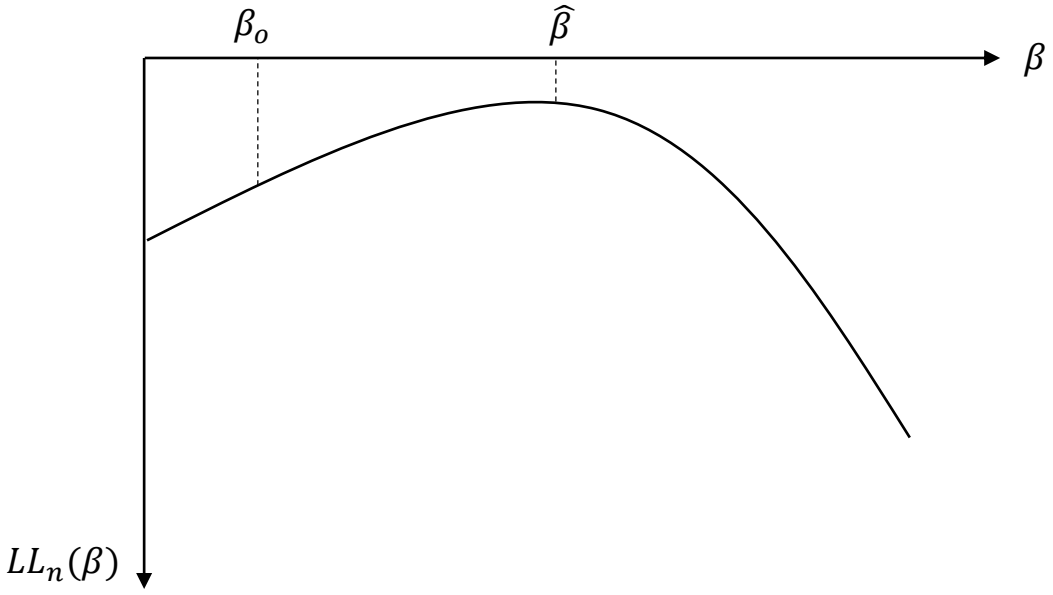
$$LL_n(\beta_1, \beta_2, \dots \beta_K) = \sum_{n=1}^{N} [y_{in} log P_n(i) + y_{jn} log P_n(j)]$$

$$= \sum_{n=1}^{N} [y_{in} log P_n(i) + (1 - y_{in}) log(1 - P_n(i))]$$

Maximum likelihood estimation of parameters

We can then obtain maximum likelihood estimates by differentiating with respect to each β, and setting the partial derivatives to equal 0 (First order Condition)

$$\frac{\partial LL}{\partial \widehat{\beta_k}} = \sum_{n=1}^{N} \left\{ y_{in} \frac{\partial P_n(i)/\partial \widehat{\beta_k}}{P_n(i)} + y_{jn} \frac{\partial P_n(j)/\partial \widehat{\beta_k}}{P_n(j)} \right\} = 0, for\ k = 1, \dots, K$$



Maximum likelihood estimate (Adapter from Train(2003))

If the likelihood function is globally concave, and a solution to the FOC exists it is unique. To prove this, the matrix of the second derivatives $\nabla^2 LL$ **(Hessian Matrix)** must be **negative semi-definite** for all values of **β**.

A **negative semi-definite** matrix is defined as such if:

$$f(x) = x'Ax \leq 0 \text{ (Quadratic form).}$$

where A is an n x n matrix (In this case our Hessian) and x a vector of values (In this case a vector of first derivatives of $LL(\beta)$evaluated at current values of β, β_o)

Maximum likelihood estimation of parameters: Binary Logit Model

The Log-likelihood function is

$$LL = \sum_{n=1}^{N} log f(x_n|\beta)$$

$$= \sum_{n=1}^{N} \left\{ y_{in} \log \left(\frac{e^{\beta' x_{in}}}{e^{\beta' x_{in}} + e^{\beta' x_{jn}}} \right) + y_{jn} \log \left(\frac{e^{\beta' x_{jn}}}{e^{\beta' x_{jn}} + e^{\beta' x_{in}}} \right) \right\}$$

Maximum likelihood estimation of parameters: Binary Logit Model

The FOC is defined as

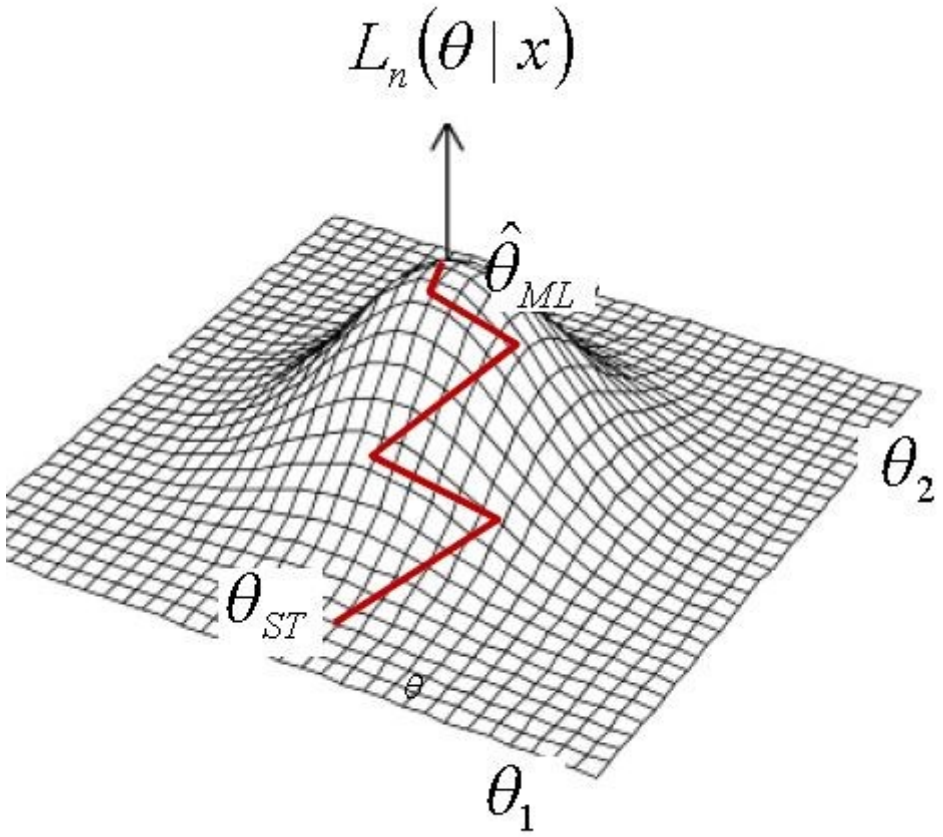
$$\frac{\partial LL}{\partial \widehat{\beta_k}} = \sum_{n=1}^{N} [y_{in} - P_n(i)][x_{ink} - x_{jnk}] = 0, k = 1, \dots, K$$

While the second derivatives can be solved as

$$\frac{\partial^2 LL}{\partial \beta \partial l} = -\sum_{n=1}^{N} [P_n(i)(1 - P_n(i))][x_{ink} - x_{jnk}][x_{inl} - x_{jnl}]$$

Maximum likelihood estimation of parameters: Algorithms

- Iterative procedures are used to estimate the ML
  - Newton-Rapshon (NR) Algorithm
  - Berdnt-Hall-Hall-Hausman (BHHH) Algorithm
  - Davidson-Fletcher-Powell (DFP) Algorithm
  - Broyden-Fletcher-Goldfarb-Shanno (BFGS) Algoritm



A concrete example: Binary Logit

Variable name	Coefficient	Standard error	t statistic
Auto constant	1.45	0.393	3.70
In-vehicle time (min)	-0.0089	0.0063	-1.42
Out-of-vehicle time (min)	-0.0308	0.0106	-2.90
Auto out-of-pocket cost (c)	-0.0115	0.0026	-4.39
Transit fare	-0.0070	0.0038	-1.87
Auto ownership (specific to auto mode)	-0.770	0.213	3.16
Downtown workplace (specific to auto mode)	-0.561	0.306	-1.84
Number of observations	1476		
Number of cases	1476		
LL(0)	-1023		
LL(β)	-347.4		
-2[LL(0)-LL(β)]	1371		
ρ^2	0.660		
$\bar{\rho}^2$	0.654		

← Magnitudes are not directly interpretable

We can only interpret the effect direction

Or to calculate utilities, and choice probabilities

To make some sense of these parameters we must calculate elasticities or marginal effects

← Log-Likelihood when all parameters are 0

← Maximum Log-Likelihood

← Test of null hypothesis that all parameters are jointly zero. χ^2 distributed

← Informal goodness-of-fit measure : $1 - (LL(\beta)/LL(0)$

← Informal goodness-of-fit measure: $1 - (LL(\beta) - K)/LL(0)$

Adapted from Ben-Akiva and Lerman (1984)

The Multinomial Logit Model

- The choice set C consists of more than two alternatives

$$P(i) = \Pr(U_{in} > U_{jn}, \forall j \in C_n, j \neq i)$$
$$P(i) = \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \forall j \in C_n, j \neq i)$$
$$= \Pr\left(\varepsilon_{jn} \leq V_{in} - V_{jn} + \varepsilon_{in}, \forall j \in C_n, j \neq i\right)$$

- We can formulate the MNL as a binary problem, so that

$$P(i) = Pr\left[V_{in} + \varepsilon_{in} \geq \max_{j \in C_n, j \neq i} (V_{jn} + \varepsilon_{jn})\right]$$

- To estimate the model we need an assumption of the **joint distribution of disturbances** $f(\varepsilon_{1n}, \varepsilon_{2n}, \varepsilon_{3n}, \dots, \varepsilon_{J_n n})$

The Multinomial Logit Model

- Error distribution assumptions:
  - Independently and identically distributed (I.I.D.)
  - Gumbel-distributed with location parameter η(usually set at 0) scale parameter μ>0 (usually set at 1)

- Under these assumptions we can derivate the MNL

$$P(i) = Pr\left[V_{in} + \varepsilon_{in} \geq \max_{j \in C_n, j \neq i} (V_{jn} + \varepsilon_{jn})\right]$$

$$P(i) = Pr[V_{in} + \varepsilon_{in} \geq V_n^* + \varepsilon_n^*] \longleftarrow (V_n^* + \varepsilon_n^*) \text{ is gumbel distributed with parameters } \left(\frac{1}{\mu} ln \sum_{j=1}^{J} \exp(\mu V_{jn}), \mu\right)$$

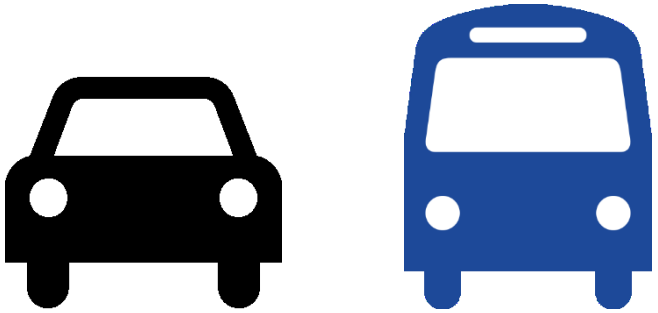
$$P(i) = Pr\left[(V_{jn}^* + \varepsilon_{jn}^*) - (V_{in} + \varepsilon_{in}) \leq 0\right] \longleftarrow \text{The difference between two Gumbel-distributed variables is Logistic-distributed}$$

$$P(i) = \frac{1}{1 + \exp\left(-\mu(V_n^* - V_{in})\right)} = \frac{\exp(\mu V_{in})}{\sum_{j \in C} \exp(\mu V_{jn})}$$

MNL: The Independence of Irrelevant Alternatives Property

For a specific individual, the ratio of the choice probabilities (Odds Ratio) of any two alternatives is unaffected by the systematic utilities of any other alternatives.

Consider a commute mode choice model where individual choose either mode with equal probabilities:



0.50 0.50

Consider then that we add a new mode (exactly the same as the other bus, but this one is red) is added. What are the choice probabilities?

To preserve the Odds Ratio, probabilities should be:

0.33 0.33 0.33

In reality however, we expect them to be:

0.50 0.25 0.25

The validity of the choice axiom only applies to choice sets with distinct alternatives.

MNL: Logit Elasticities (Point elasticities)

- **Direct elasticity:** measures the **percentage change in the probability** of choosing a particular alternative in the choice set with respect to a given **percentage change** in an attribute of that same alternative.

$$E_{x_{ink}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} \cdot \frac{x_{ink}}{P_n(i)} = [1 - P_n(i)] x_{ink} \beta_k$$

- **Cross-elasticity:** measures the **percentage change in the probability** of choosing a particular alternative in the choice set with respect to a given **percentage change** in a competing alternative.

$$E_{x_{jnk}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} \cdot \frac{x_{jnk}}{P_n(i)} = -P_n(j) x_{jnk} \beta_k$$

← Because of IIA, cross-elasticities are uniform across all alternatives

Definition following Louviere, Hensher, and Swait (2000)

MNL: Logit Elasticities (Point elasticities)

- The elasticities shown before are **individual elasticities (Disaggregate)**
- To calculate sample (aggregate) elasticities we use the **probability weighted sample enumeration** method:

$$E_{x_{ink}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) E_{x_{ink}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

Sample direct elasticity

$$E_{x_{jnk}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) E_{x_{jnk}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

Sample cross-elasticity

Where $\overline{P(i)}$ is the aggregate choice probability of alternative I, and $\hat{P}_{in}(i)$ is an estimated choice probability

- Uniform cross-elasticities do not necessarily hold at the aggregate level
- Also note that elasticities for dummy variables are **meaningless!**

MNL: Logit Marginal Effects

- **Direct marginal effects:** measures the **change in the probability** (absolute change) of choosing a particular alternative in the choice set with respect to a **unit change** in an attribute of that same alternative.

$$M_{x_{ink}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} = [1 - P_n(i)]\beta_k$$

- **Cross-marginal effects:** measures the **change in the probability** (absolute change) of choosing a particular alternative in the choice set with respect to a **unit change** in a competing alternative.

$$M_{x_{jnk}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} = -P_n(j)\beta_k$$

Definition following Louviere, Hensher, and Swait (2000)

MNL: Logit Marginal Effects

- We can also calculate sample (aggregate) marginal effects we using e the **probability weighted sample enumeration** method:

$$M_{x_{ink}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) M_{x_{ink}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

Sample direct marginal effect

$$M_{x_{jnk}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) M_{x_{jnk}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

Sample cross-marginal effect

Where $\overline{P(i)}$ is the aggregate choice probability of alternative I, and $\hat{P}_{in}(i)$ is an estimated choice probability

- Marginal effects for dummy variables **do make sense** as we are talking about unit changes!

Incremental Logit for prediction

- An alternative approach to using elasticities or marginal effects for prediction
- Prediction of changes in behavior based on existing choice probabilities

$$P'(i) = \frac{\exp(V_{in} + \Delta V_{in})}{\sum_{j \in C} \exp\left(V_{jn} + \Delta V_{jn}\right)}; \qquad where\ \Delta V_{in} = \sum_{k=1}^{K} \beta_k \Delta x_{ink}$$

Δx_{ink} is a marginal change in the k^{th} independent variable for alternative i and individual n

- In fact, for linear-in-parameter models we need not calculate the utilities again

$$P'(i) = \frac{\exp(V_{in} + \Delta V_{in})}{\sum_{j \in C} \exp\left(V_{jn} + \Delta_{jn}\right)} = \frac{\mathrm{P}(i)\exp(\Delta V_{in})}{\sum_{j \in C} P(j) \exp\left(\Delta V_{jn}\right)}$$

A concrete example: Multinomial Logit Model

Variable name	Coefficient	Standard error	t statistic
Drive-alone constant	-3.24	0.472	-6.90
Shared-ride constant	-2.24	0.400	-5.6
In-vehicle time (min)	-0.015	0.0057	-2.7
Out-of-vehicle time/distance (min/mile)	-0.160	0.0392	-4.1
Cost (c)/annual income ($/year)	-28.8	12.7	-2.3
Car to driver ratio (drive-alone)	3.99	0.396	10.1
Car to driver ratio (shared-ride)	1.62	0.305	5.3
Downtown workplace dummy (drive-alone)	-0.854	0.311	-2.8
Downtown workplace dummy (shared-ride)	-4.04	0.297	-1.4
Disposable income ($/yr) (drive alone, shared-ride)	0.00007	0.00002	3.5
Primary worker dummy (drive-alone)	0.890	0.186	4.8
Government worker dummy (shared ride)	0.287	0.161	1.8
Number of workers (shared ride)	0.0983	0.0954	1.0
Employment distance x Distance (shared ride)	0.00063	0.00047	1.3

Adapted from Ben-Akiva and Lerman (1984)

Choice set:

1. Driving alone
2. Sharing a ride
3. Transit (Bus)

Number of observations	1114
Number of cases	2932
LL(0)	-1054
LL(β)	-727.4
-2[LL(0)-LL(β)]	653.2
ρ^2	0.309
$\bar{\rho}^2$	0.297

A concrete example: Multinomial Logit Model

Writing down the utility functions

$$V_{Drivealone} = \beta_d + \beta_{ivt} In.vehicle.time + \beta_{ovt} Out.of.vehicle.time + \beta_{cost} Cost + \beta_{ctdr.d} Car.to.driver.Ratio + \beta_{dwnt.d} Downtown.workplace + \beta_{disnc} Disposable.income + \beta_{pworker} Primary.worker$$

$$V_{Sharedride} = \beta_s + \beta_{ivt} In.vehicle.time + \beta_{ovt} Out.of.vehicle.time + \beta_{cost} Cost + \beta_{ctdr.s} Car.to.driver.Ratio + \beta_{dwnt.s} Downtown.workplace + \beta_{disnc} Disposable.income + \beta_{Govwrk} Goverment.worker + \beta_{nowkrs} Number.of.workers + \beta_{densdist} Employment.density.distance$$

$$V_{Bus} = \beta_{ivt} In.vehicle.time + \beta_{ovt} Out.of.vehicle.time + \beta_{cost} Cost$$

A concrete example: Multinomial Logit Model

Writing down the utility functions

$$V_{Drivealone} = -3.24 - 0.015 \cdot In.vehicle.time - 0.160 \cdot Out.of.vehicle.time - 28.8 \cdot Cost + 3.99 \cdot Car.to.driver.Ratio - 0.854 \cdot Downtown.workplace + 0.00007 \cdot Disposable.income + 0.890 \cdot Primary.worker$$

$$V_{Sharedride} = -2.24 - 0.015 \cdot In.vehicle.time - 0.160 \cdot Out.of.vehicle.time - 28.8 \cdot Cost + 1.62 \cdot Car.to.driver.Ratio - 4.04 \cdot Downtown.workplace + 0.00007 \cdot Disposable.income + 0.287 \cdot Goverment.worker + 0.0983 \cdot Number.of.workers + 0.00063 \cdot Employment.density.distance$$

$$V_{Bus} = -0.015 \cdot In.vehicle.time \pm 0.160 \cdot Out.of.vehicle.time \pm 28.8 \cdot Cost$$

A concrete example: Multinomial Logit Model

Using the incremental logit for prediction

Assume that for individual n, the choice probabilities are as follow:

- *Driver alone: 0.65* *Share ride: 0.31* *Transit (Bus): 0.04*

$$P(drive\ alone) = \frac{\exp(\mu V_{in})}{\sum_{j \in C} \exp(\mu V_{jn})} = \frac{0.1460}{0.1460 + 0.0689 + 0.010} = \frac{0.1460}{0.2249} = \mathbf{0.65}$$

Assume that a road expansion project reduces individual n's:

- *In-vehicle time by 15 minutes (from 25 to 10 minutes) when driving alone*
- *In-vehicle time by 5 minutes (from 35 to 30 minutes) when sharing a ride (needs to drive wife to work first)*

$$P'(drive\ alone) = \frac{\mathrm{P}(i)\exp(\Delta V_{in})}{\sum_{j \in C} P(j) \exp(\Delta V_{jn})} = \frac{0.65 \cdot \exp(0.025)}{0.31 \cdot \exp(0.075) + 0.04 \cdot \exp(0)} = \mathbf{0.68}$$

A concrete example: Multinomial Logit Model

Using the incremental logit for prediction

The key point is that we can use changes in individual behavior to calculate aggregate values (we will discuss that in further lectures) and evaluate how a policy implementation affects travel behavior (in this case, mode choice).

Aggregate forecasting techniques

- Why is it important?
  - So far we have dealt only with **individual probabilities.**
  - But we are interested in **aggregate forecasts in order to make planning decisions.**

- The first issue to address:
  - **Define the population of interest *T*:**
    - *All the residents of the city of interest?*
    - *A specific segment? (i.e. income group, racial group, etc.)*

  - Generally, we can **use existing data sources** such as the **national census** to estimate the size of T.

  - **Define:**
    - N_T : *the number of decision makers*
    - $P(i|\boldsymbol{x}_n)$: *the probability of individual n choosing alternative i given attributes* $\boldsymbol{x}_n$

Aggregate forecasting techniques

Attributes of the MNL mode choice model presented in last class

Variable name
In-vehicle time (min)
Out-of-vehicle time/distance (min/mile)
Cost (c)/annual income ($/year)
Car to driver ratio (drive-alone)
Car to driver ratio (shared-ride)
Downtown workplace dummy (drive-alone)
Downtown workplace dummy (shared-ride)
Disposable income ($/yr) (drive alone, shared-ride)
Primary worker dummy (drive-alone)
Government worker dummy (shared ride)
Number of workers (shared ride)
Employment distance x Distance (shared ride)

Provided we know the values of $\mathbf{x_n}$ for all n, then the expected number of individuals in T choosing i **(that is, the expected value of the aggregate number of individuals)** is**:**

$$N_T(i) = \sum_{n=1}^{N_T} P(i|\boldsymbol{x_n})$$

More conveniently, we can express this equation as ratio (market share):

$$W(i) = \frac{1}{N_T}\sum_{n=1}^{N_T} P(i|\boldsymbol{x_n}) = \mathbb{E}[P(i|\boldsymbol{x_n})]$$

When $\mathbf{x_n}$ is continuous in T, W is defined as the following integral

$$W(i) = \int_{\boldsymbol{x}} P(i|\boldsymbol{x})p(\boldsymbol{x})\boldsymbol{dx}$$

$p(\boldsymbol{x})$ is unknown, and evaluating this integral might be computationally burdensome.

Aggregate forecasting techniques

In short, we require methods that reduce the required data and computational needs to predict aggregate shares.

- General approaches to aggregate forecasting (Koppelman, 1975):
  - **Average individual**
  - **Classification**
  - Statistical differentials (inappropriate in very heterogeneous populations)
  - Explicit integration (too difficult to apply in multinomial cases)
  - **Sample enumeration**

We will focus the **Average individual approach** as a departure point to elaborate on the most frequently used methods empirically; **namely classification and sample enumeration**.

Aggregate forecasting techniques

① Average Individual

*Constructs an **"average individual"** for the population and uses the choice probability for the average individual as an approximation of W(i).*

More specifically:

- Define $\bar{\mathbf{x}}$ as the mean of $\mathrm{p}(\mathbf{x})$
- Our approximation of $\mathrm{W}(\mathrm{i})$ is thus

$$W(i) = \mathbb{E}[P(i|\boldsymbol{x})] \cong P(i|\bar{\boldsymbol{x}}_{\boldsymbol{n}})$$

A simple example: Assume that T=3, and we know the true values of $\mathbf{x_n}$, then the true value of $\mathrm{W}(\mathrm{i})$ is :

$$W(i) = \frac{P(i|x_1) + P(i|x_2) + P(i|x_3)}{3}$$

- Only when the model is linear over the range of x, the aggregation error Δ will be zero.
- the aggregation error Δ increases **as the variance of p(x) increases.**

Aggregate forecasting techniques

② Classification

Divides the population into G nearly homogenous subgroups, and uses the choice probabilities of the average individual within each subgroup. $N_T(i)$ *is estimated as the weighted sum of each subgroups' forecasts.*

More specifically:

- Partition T into G mutually exclusive, collectively exhaustive subgroups.
- For each subgroup, choose a representative value $\widetilde{\mathrm{x_g}}$.
- Approximate $W(i)$ as

$$W(i) = \mathbb{E}[P(i|\boldsymbol{x})] \cong \sum_{g=1}^{G} \frac{N_g}{N_T} P(i|\tilde{\boldsymbol{x}}_{\boldsymbol{g}})$$

Aggregate forecasting techniques

② Classification

$$W(i) = \mathbb{E}[P(i|\boldsymbol{x})] \cong \sum_{g=1}^{G} \frac{N_g}{N_T} P(i|\tilde{\boldsymbol{x}}_{\boldsymbol{g}})$$

- As the number of subgroups increases, so does estimate accuracy, but it comes at higher data and computational requirements.

- Most of the times, it is unfeasible to classify on every dimension in **x**. Hence we need good judgment in deciding sub-classification variables.

- Some insights regarding sub-classification:
  - *Select **a small number** of independent variables which:*
    - *Have a **large effect on the systematic utility** of at least one alternative*
    - *Have a **wide distribution** across the population*
  - *All else equal, **avoid disproportionately small classes****

Aggregate forecasting techniques

③ Sample enumeration

Uses a sample to represent the entire population.

- **When using random sampling**

$$\widehat{W}(i) = \frac{1}{N_S}\sum_{n=1}^{N_S} P(i|\boldsymbol{x_n})$$

- **When using nonrandom sampling (i.e. Stratified sampling)**

$$\widehat{W}(i) = \sum_{g=1}^{G}\left(\frac{N_g}{N_T}\right)\frac{1}{N_{Sg}}\sum_{n=1}^{N_{Sg}} P(i|\boldsymbol{x_n})$$

Aggregate forecasting techniques

③ Sample enumeration

- Predicted aggregate shares are estimates, and as such are subject to sampling error.
  - When choice probabilities or samples are small, sampling error might be a large fraction of W(i).

- Sample enumeration makes it easy to produce forecasts for different socio-economic groups, provided sample sizes are large enough (as we saw in the previous lecture).

Relevant statistical tests

- To some extent, **modeling is an "art"** as much as is a science.
- **We cannot rely exclusively on goodness-of-fit statistics.**
- **Several model specifications might fit** the data as well.
- Good fitting models can **still result in erroneous predictions.**
- **Theory and informal judgment** play an important role.

Relevant statistical tests

① Testing coefficient estimates

- Are signs consistent with our expectations? ←Informal test

Variable name	Coefficient	Standard error	t statistic
…			
4. In-vehicle time (min)	-0.015	0.0057	-2.7
5. Cost (c)/annual income ($/year)	-28.8	12.7	-2.3
6. Car to driver ratio (drive-alone)	3.99	0.396	10.1
7. Car to driver ratio (shared-ride)	3.88	0.376	10.3
…			

A positive sign for cost should ring some alarms

Are this parameters statistically different from one another?

- Are the parameters statistically significant? ← Asymptotic t Test
  - Same as in linear regression, but only valid for **large sample sizes**
- Asymptotic t Test for linear relationships among parameters

$$t = \frac{\hat{\beta}_6 - \hat{\beta}_7}{\sqrt{var(\hat{\beta}_6 - \hat{\beta}_7)}}; \qquad where\ \ H_o: \beta_6 = \beta_7$$

Relevant statistical tests

② The likelihood ratio test: $-2\left(LL(0) - LL(\hat{\boldsymbol{\beta}})\right)$

- $\mathrm{H_o}: \beta_1 = \beta_2 = \cdots = \beta_K = 0$ ←Similar to the F-test in OLS regression
- X^2 distributed with K degrees of freedom
- Not very useful. $\mathrm{H_o}$ is almost always rejected!
- **More useful applications of the likelihood ratio test:**
  - ① Compare against a constant only model: $-2\left(LL(\boldsymbol{C}) - LL(\hat{\boldsymbol{\beta}})\right)$

    Where, $LL(\boldsymbol{C}) = \sum_{i=1}^{J} N_i \ln\left(\frac{N_i}{N}\right)$, X^2 distributed with $K - J + 1$ degrees of freedom.
  - ② Comparing nested models: $-2\left(LL(\hat{\boldsymbol{\beta}}_r) - LL(\hat{\boldsymbol{\beta}}_u)\right)$

    Where $LL(\hat{\boldsymbol{\beta}}_r)$ is the Log-likelihood of the restricted model, $LL(\hat{\boldsymbol{\beta}}_r)$ the log-likelihood of the unrestricted model. (Test of linear relations, generic parameters etc)

    X^2 distributed with $(K_u - K_r)$ degrees of freedom.

Relevant statistical tests

③ Goodness of fit test:

$$\rho^2 = 1 - \frac{LL(\hat{\boldsymbol{\beta}})}{LL(0)}$$ ← Used in a similar manner to R2 in OLS regression.

$$\bar{\rho}^2 = 1 - \frac{LL(\hat{\boldsymbol{\beta}}) - K}{LL(0)}$$ ← Favors more parsimonious specifications (unless newly added variables are very significant).

- All else equal, specifications with higher goodness of fit values should be selected.
- Can be used to test non-nested hypotheses of discrete choice models.
- Most useful when comparing models estimated using the same dataset.

Relevant statistical tests

③ Testing the IIA assumption

- So far we have assumed that our model structure is correct. But is this so?
- Recall that the IIA assumption is critical for the validity of the logit model structure
- Two general tests:
  - The Haussmann and McFadden Test (1984)
  - Approximate likelihood ratio test (Small and Hsiao (1982))
- In general, the tests consist on **comparing logit models estimated with subset of alternatives** from the universal choice set.
- If the IIA assumption holds for the full choice set, **the same model structure (logit) should apply to model with a subset of alternatives** (the restricted model).

Relevant statistical tests

③ Testing the IIA assumption: Haussmann and McFadden Test (Most frequently used)

$$q = \left[\hat{\boldsymbol{\beta}}_{\boldsymbol{u}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{r}}\right]' \left[\Sigma_r - \Sigma_u\right]^{-1} \left[\hat{\boldsymbol{\beta}}_{\boldsymbol{u}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{r}}\right]$$

*Note that constant terms are not included in the calculation of q.

$\hat{\beta}_u$ is a column vector of parameter estimates for the unrestricted model

$\hat{\beta}_r$ is a column vector of parameter estimates for the restricted model

Σ_u is the variance-covariance matrix of the unrestricted model

Σ_r is the variance-covariance matrix of the restricted model

- $H_o: \hat{\boldsymbol{\beta}}_{\boldsymbol{u}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{r}}$ ← **Note that failing to reject the Null Hypotheses means that the IIA Holds!**

- q is asymptotically X^2 distributed with K_r degrees of freedom (The dimension of the restricted model)

Relevant statistical tests

③ Testing the IIA assumption: Approximate likelihood ratio test

$$q = \frac{1}{1 - N_1/\alpha N}\{-2[LL_r(\hat{\beta}_u) - LL_r(\hat{\beta}_r)\}$$

$LL_r(\hat{\beta}_u)$is the log-likelihood values for the unrestricted model calculated on the restricted sample

$LL_r(\hat{\beta}_r)$is the log-likelihood values for the restricted model calculated on the restricted sample

N is the number of observations in the unrestricted choice set estimation

N_1 is the number of observations in the restricted choice set estimation

$\alpha \geq 1$ is a scalar (For screening purposes usually assumed as 1)

- $H_o: \widehat{\boldsymbol{\beta}}_{\boldsymbol{u}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{r}}$ ← **Note that failing to reject the Null Hypotheses means that the IIA Holds!**
- q is asymptotically X^2 distributed with K_r degrees of freedom (The dimension of the restricted model) **if the difference between the covariance matrices differs at most by a scalar multiple.**

Relevant statistical tests

④ Testing for taste variations

- So far we have assumed that the parameters are the same for all members of the population. (i.e. the magnitude of the effects are the same) **How can we test if this is in fact true?**
    - ① Allow for random taste variation in coefficients (Random parameter models)
    - ② Market segmentation

Relevant statistical tests

④ Testing for taste variations : Market segmentation

*Include **socio-demographic characteristics** to account for unobservable taste variations.*

More specifically:

- **Classify the sample data** into socio-economic groups (e.g. Income groups, car ownership, etc.)
- **Estimate separate models** (same specification across markets) for each sub-group **and a pooled model with the full dataset.**
- Use the likelihood ratio test where $\mathrm{H_o}$: $\boldsymbol{\beta^1 = \beta^2 = \cdots = \beta^G}$

$$-2\left[LL_N(\hat{\beta}_{full}) - \sum_{g=1}^{G} LL_{N_g}(\hat{\beta}^g)\right]$$

X^2 distributed with $\sum_{g=1}^{G} K_g - K$ degrees of freedom

$LL_N(\hat{\beta}_{full})$ is the log-likelihood of the pooled model (non-segmented)

$LL_{N_g}(\hat{\beta}^g)$ is the log-likelihood of the model estimated with the g^{th} data subset

Relevant statistical tests

④ Testing for taste variations :

$$-2\left[LL_N(\hat{\beta}_{full}) - \sum_{g=1}^{G} LL_{N_g}(\hat{\beta}^g)\right] = -2[-820.3 + 803.7] = 33.2$$

Degrees of freedom: 12 $X^2_{0.05} = 21.0$

We thus reject the null hypothesis that $\boldsymbol{\beta^1 = \beta^2}$

Individual coefficients can also be compared across Segments:

$$t = \frac{\widehat{\beta^1}_k - \widehat{\beta^2}_k}{\sqrt{var(\widehat{\beta^1}_k) + (\widehat{\beta^2}_k)}}; \qquad where \;\; H_o : \widehat{\beta^1}_k = \widehat{\beta^2}_k$$

Note that it is certainly possible that:

- All t tests are insignificant despite the joint likelihood being significant.
- The joint test does not reject the null hypothesis but some coefficients might be significantly different.

MNL Model segmented by auto ownership levels

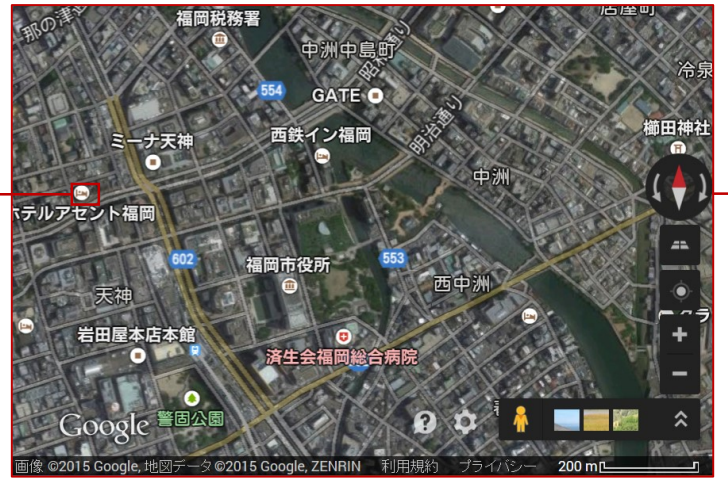
Variable Name	Segment 1 Auto Ownership (0 or 1)	Segment 2 Auto Ownership (2+)
Drive alone (DA) constant	-2.660	-3.240
	(-5.846)	(-2.436)
Shared ride (SR) constant	-1.140	-2.980
	(-3.826)	(-2.463)
Rount-trip travel time (min)	0.028	-0.049
	(3.500)	(-2.455)
Round-trip out-of-vehicle time (min)/ one-way distance (0.01 mile)	-14.700	-14.500
	(-2.341)	(-1.295)
Cars/workers in household (DA specific)	-35.300	-35.400
	(-1.929)	(1.009)
Cars/workers in household (SR specific)	4.260	3.560
	(9.861)	(3.849)
Downtown workplace dummy (DA specific)	1.400	2.590
	(4.106)	(2.776)
Downtown workplace dummy (SR specific)	-0.605	-1.130
	(-1.644)	(-1.865)
Disposable household income (DA specific)	-0.446	-0.636
	(-1.502)	(-1.102)
Disposable household income (SR specific)	0.000	0.001
	(1.335)	(24.901)
Government worker dummy (SR specific)	0.687	0.063
	(3.435)	(0.251)
Observations per segment	623	513
$LL_{N_g}(\hat{\beta}^g)$	-502.600	-301.100
Total observations = 1,136		
$LL_N(\hat{\beta}_{full})$ = -820.3		

Adapted from Ben-Akiva and Lerman (1984)

Aggregation and sampling of alternatives

- So far we have discussed choice models using clearly defined mode alternatives.
- What happens when we consider spatial choices? (e.g. residential location, destination choice?) How do we define the choice set?

Lets's think about residential location choice.



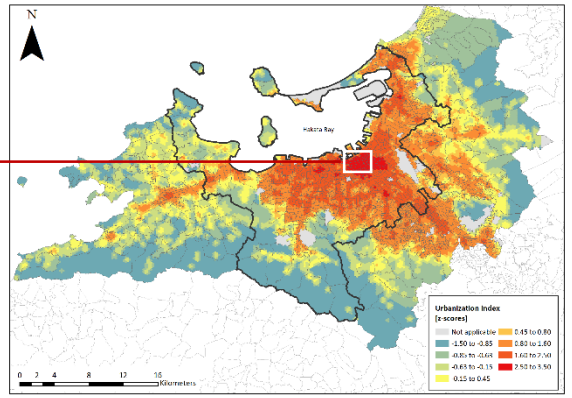


Image source: trendy.nikkeibp.co.jp

Aggregation and sampling of alternatives

Why aggregation and sampling?

- In most spatial choices, we cannot identify the actual alternatives the individual chooses from.
- Even if we could, the number of alternatives is just too large.
- Also, data on the attractiveness of each alternative is usually aggregated to some extent.

Aggregation and sampling of alternatives

Elemental alternatives:

Actual alternatives an individual chooses from. Where $\mathrm{p_n(l)}$ is the probability of individual n choosing alternative $l \in L$.

The utility of elemental alternative l is thus:

$$U_{ln} = V_{ln} + \varepsilon_{ln}$$

Aggregate alternatives:

By dividing the universal set of elemental alternatives into non-overlapping subsets we get:

$$L_i \subseteq L, i = 1, \dots, J$$

Thus the choice probability of choosing aggregate alternative i is:

$$P_n(i) = \sum_{l \in L_i} p_n(l), \qquad i = 1, \dots, J$$

The utility of aggregate alternative i is then:

$$U_{in} = \max_{l \in L_i}(V_{ln} + \varepsilon_{ln}), \qquad i = 1, \dots, J$$

It can be shown that if an aggregate alternative contains a large number of elemental alternatives, and if the utilities of these alternatives are IID, then **the distribution of the aggregate alternative is Gumbel distributed.**

Aggregation and sampling of alternatives

A logit model with aggregate alternatives:

$$P(i) = \frac{\exp(\mu^* V_{in})}{\sum_{j=1}^{J} \exp(\mu^* V_{jn})} = \frac{\exp(\mu^* \bar{V}_{in} + \mu' lnM_i + \mu' lnB_{in})}{\sum_{j=1}^{J} \exp(\mu^* \bar{V}_{in} + \mu' lnM_i + \mu' lnB_{in})}$$

$\bar{V}_{in}$ is the average utility of the elemental alternatives in aggregate alternative i

μ^* is a positive scale parameter, and $\mu' = \mu^*/\mu$ is the coefficient o the log of the size measure

lnM_i is the log of the measure of size of the aggregate alternative, that is, **the number of elemental alternatives that compose the aggregate.** (Think carefully about what that the elemental alternative might be!)

lnB_{in} is the log of the measure of variability of the utilities of elemental alternatives in the aggregate alternative. **This term can be omitted if all elemental alternatives have equal variance.**

Aggregation and sampling of alternatives

A logit model with aggregate alternatives:

Thus we have

$$P(i) = \frac{\exp(\mu^* V_{in})}{\sum_{j=1}^{J} \exp(\mu^* V_{jn})} = \frac{\exp(\mu^* \bar{V}_{in} + \mu' ln M_i)}{\sum_{j=1}^{J} \exp(\mu^* \bar{V}_{in} + \mu' ln M_i)}$$

What happens if we cannot estimate the number of elemental alternatives M_i ?

We can use a single proxy variable, and keep the same specification as above. In the case of residential location we could use: *Number of households, Population density, or area.*

We can use a set of proxy variables, where

$$M_i = \sum_{s=1}^{S} \beta_s x_{ins}, \qquad i = 1, \dots, J$$

Not including a size variable forces us to include J-1 ASCs in order to properly capture the size effect!

S is the number of size variables used

x_{ins} are the size variables included in the model (x_{ins} and its coefficients must be non-negative!)

Aggregation and sampling of alternatives

Disadvantages of aggregation:

- Aggregation introduces measurement error in the explanatory variables, thus reducing the model accuracy.
- The Modifiable Areal Unit Problem: The way spatial zones are aggregated or subdivided for analysis might affect the results in unpredictable ways (Fotheringham & Wong (1991).
- On the other hand, analysis using disaggregate alternatives might be prohibitively expensive computationally, and burdensome in terms of data management.

Why sampling of alternatives?

- Can mitigate some of the disadvantages of aggregation.
- Can address the computational and data management burdens while still allowing for consistent parameter estimation. (Thanks to the IIA property)

Aggregation and sampling of alternatives

Sampling of alternatives

- McFadden (1978) showed that a consistent estimator for the logit model using samples of alternatives can be estimated via a conditional probability:

Alternative-specific bias correction term

$$\pi_n(i|\boldsymbol{D}) = \frac{\exp(\mu^* V_{in} + \overbrace{ln\pi_n(\boldsymbol{D}|i)}^{\text{Alternative-specific bias correction term}})}{\sum_{j\in\boldsymbol{D}} \exp\big(\mu^* V_{jn} + \underbrace{ln\pi_n(\boldsymbol{D}|i)}_{\text{Alternative-specific bias correction term}}\big)}, \qquad i \in \boldsymbol{D}$$

$\pi_n(i|\boldsymbol{D})$ is the conditional probability for observation n of choosing alternative I given a subset of alternatives **D**.

$\pi_n(\boldsymbol{D}|i)\ and\ (\boldsymbol{D}|j)$ are the conditional probabilities of constructing for observation n a set **D**, given that the chosen alternative is i and j respectively.

Aggregation and sampling of alternatives

Sampling of alternatives: Random sampling

- Draw randomly (Without replacement) J' alternatives from all the available alternatives, excluding the chosen alternative, and then adding the chosen one, so that

$$\pi_n(\boldsymbol{D}|i) = \pi_n(\boldsymbol{D}|j) = \binom{J-1}{J'}^{-1}, \qquad \forall \mathrm{i,j} \in \boldsymbol{D}$$

- The binomial coefficient $\binom{J-1}{J'}$ determines the possible number of combinations of J items.

- Since $\ln\pi_n(\boldsymbol{D}|i) = ln\pi_n(\boldsymbol{D}|j)$ the sampling correction terms cancel out in the conditional probability equation, and **the model can be estimated using the ordinary logit model.**

$$\pi_n(i|\boldsymbol{D}) = \frac{\exp(\mu^* V_{in})}{\sum_{\boldsymbol{j}\in\boldsymbol{D}} \exp(\mu^* V_{jn})}, \qquad i \in \boldsymbol{D}$$

Aggregation and sampling of alternatives

Sampling of alternatives: Importance based sampling

- The probability of an alternative being selected depends on the likelihood of it being chosen by the decision maker.
- Based on preliminary estimates of choice probabilities, usually estimated using simpler model forms. (i.e. a gravity-type function for destination choice etc…)
- Several types of sampling:
  - *Independent Importance Sampling*
  - *Importance Sampling with Replacement*
  - *Stratified Importance Sampling*
- The conditional choice probability is given by

$$\pi_n(i|\boldsymbol{D}) = \frac{\exp(\mu^* V_{in} - lnq_{in})}{\sum_{j \in D} \exp\left(\mu^* V_{jn} - lnq_{jn}\right)}, \qquad i \in \boldsymbol{D}$$

where $q_{in}\ and\ q_{jn}$ is the estimated selection probabilities

Aggregation and sampling of alternatives

Sampling of alternatives: Importance based sampling

- ***Independent Importance Sampling:*** *Draws J-1 independent draws from the set of all alternatives excluding the chosen alternative, selecting alternative j with selection probability* q_{jn}*. Then the add the chosen alternative.*

- ***Importance Sampling with Replacement:*** *Draw a sample of size J' from the set of all alternatives, selecting alternative j with selection probability* q_{jn} *at each draw. Delete duplicate alternatives and add the chosen alternative if it was not sampled.*

- ***Stratified Importance Sampling:*** *Stratify the sample into R subsets. Assign different choice probabilities to each subset, and randomly sample (without replacement)* $\widetilde{J_{nr}}$ *draws from each strata. (For the stratum that contains the chosen alternatives draw a sample of only* $J_{nr} - 1$*, and then add the chosen alternative.*