The 17th Behavior Modeling Summer School

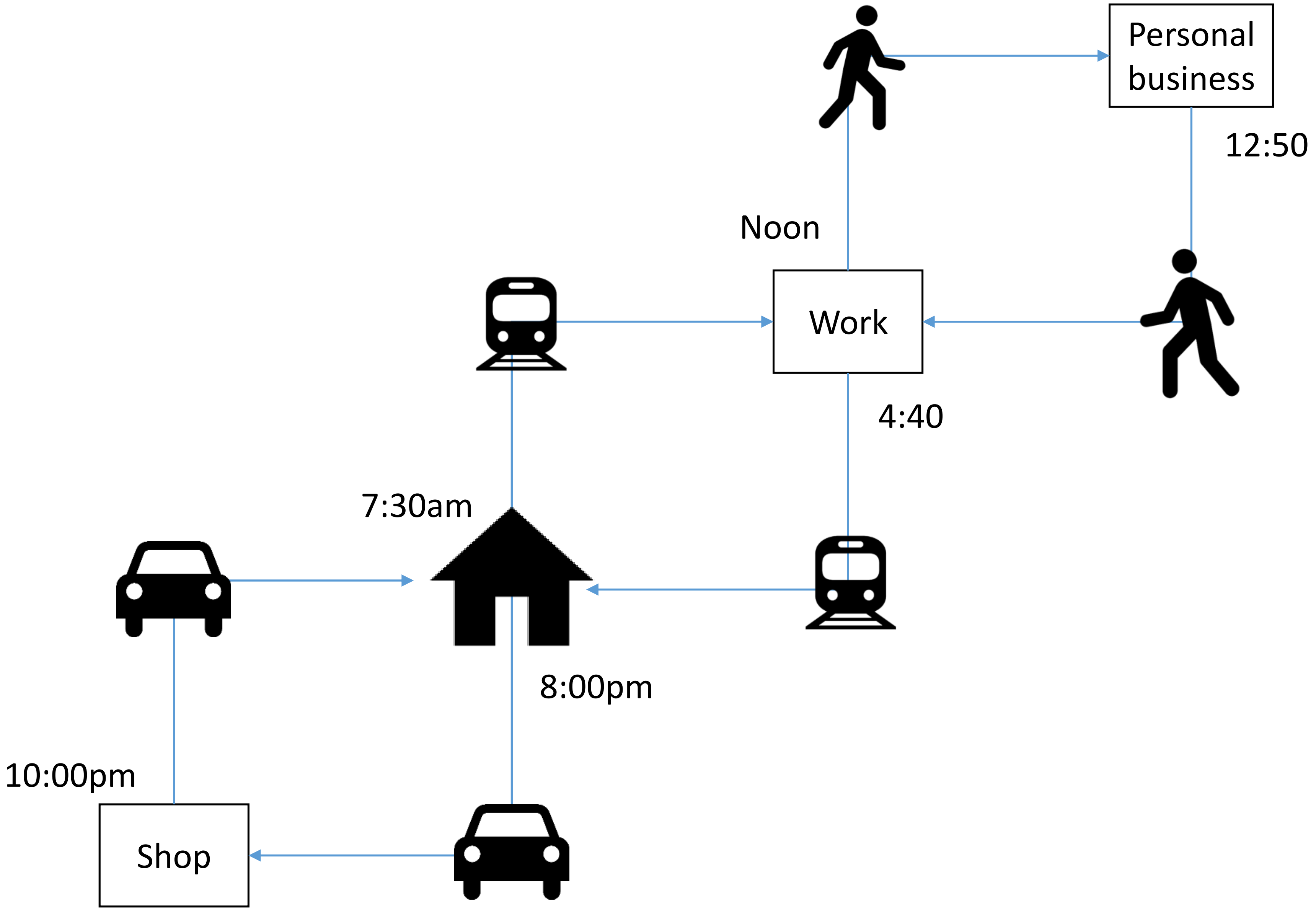
September 14-16, 2017

Introduction to Discrete Choice Models

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

THE UNIVERSITY OF TOKYO

Discrete choice theory



Bowman and Ben-akiva 1997

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

Discrete choice theory

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****)*

Discrete choice theory

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

$$U_{in} = \underbrace{V_{in}}_{\text{Observable (systematic) component}} + \underbrace{\varepsilon_{in}}_{\text{Unobservable (random) component}}$$

So that

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

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

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

Only difference in utility matters!

Where C_n is a feasible choice set for individual n

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

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

Discrete choice theory

An utility-maximization decision rule

- Specifying the utility function components

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

- Usually linear-in-parameters specification:

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

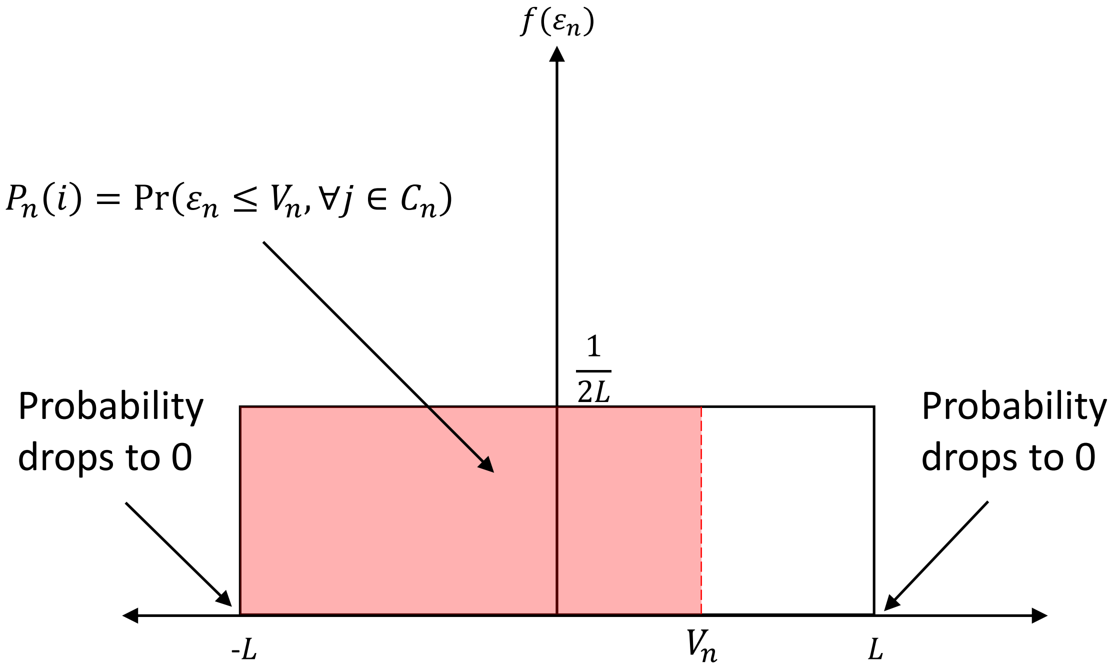
$$where\ x_{in} = f(z_{in}, S_n)$$

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

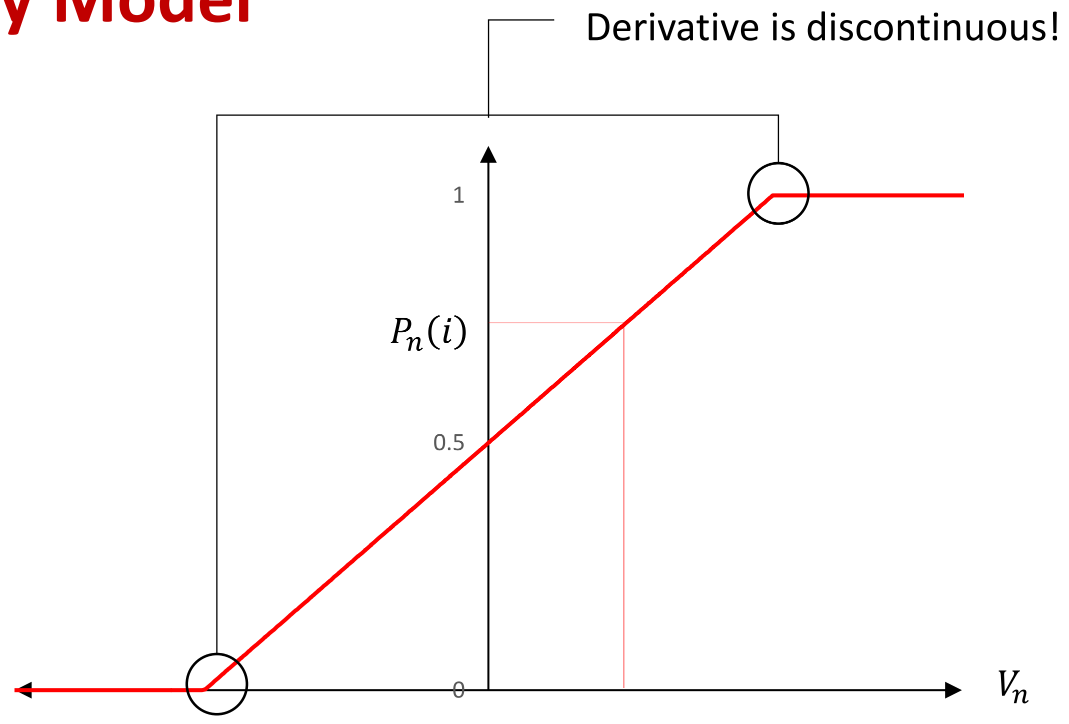
- Reflects the sources of randomness discussed earlier
- Different distributional assumptions result in different models:
  - *Normal distribution→Probit model*
  - *Gumbel distribution→Logit model*

Discrete choice theory

Binary choice models: Linear Probability Model



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



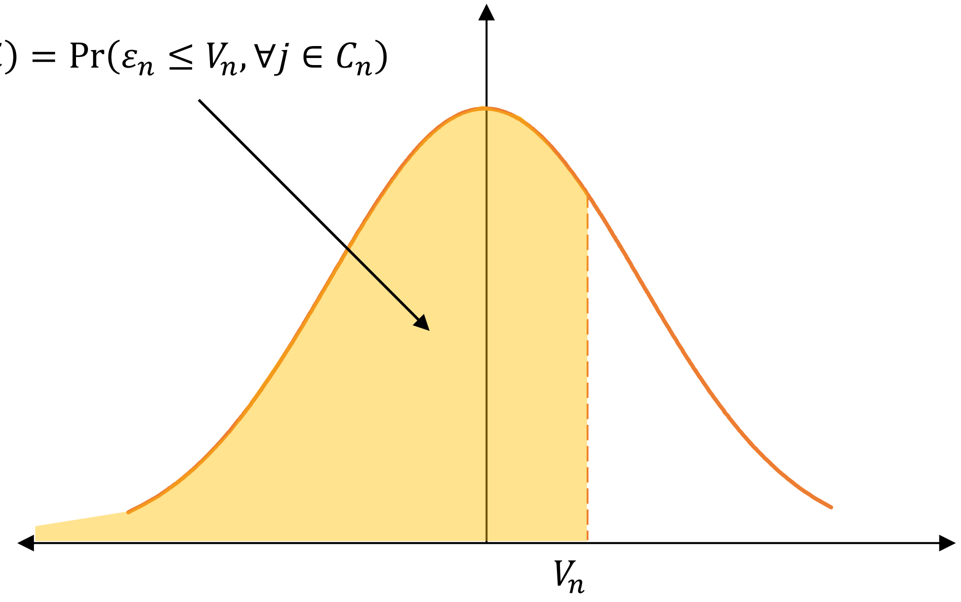
- The choice probability of i is given by the CDF of ε_n

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

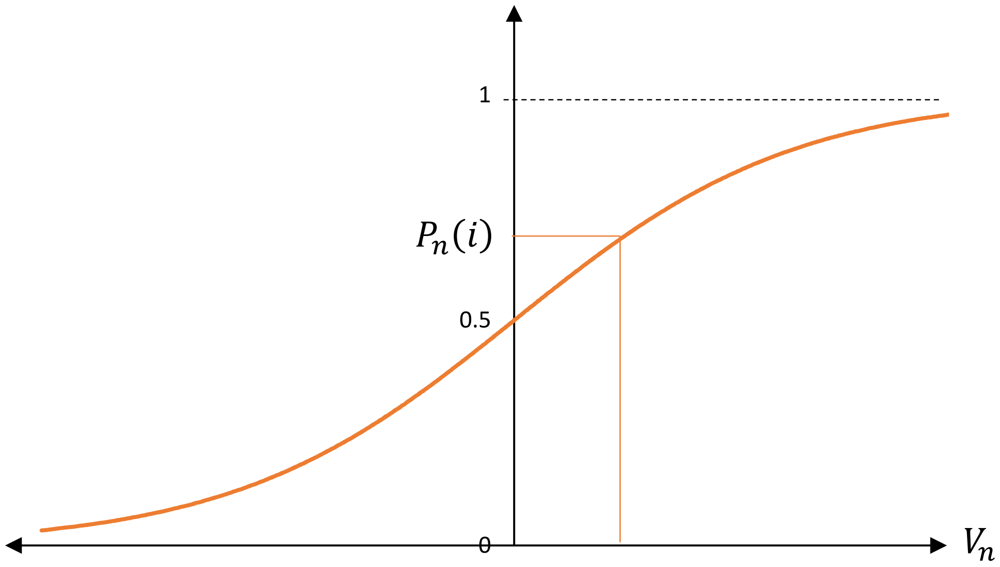
Choices with predicted probability of 0 are still chosen.

Discrete choice theory

Binary choice models: Probit Model



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



Normal distribution CDF of ε_n

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

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

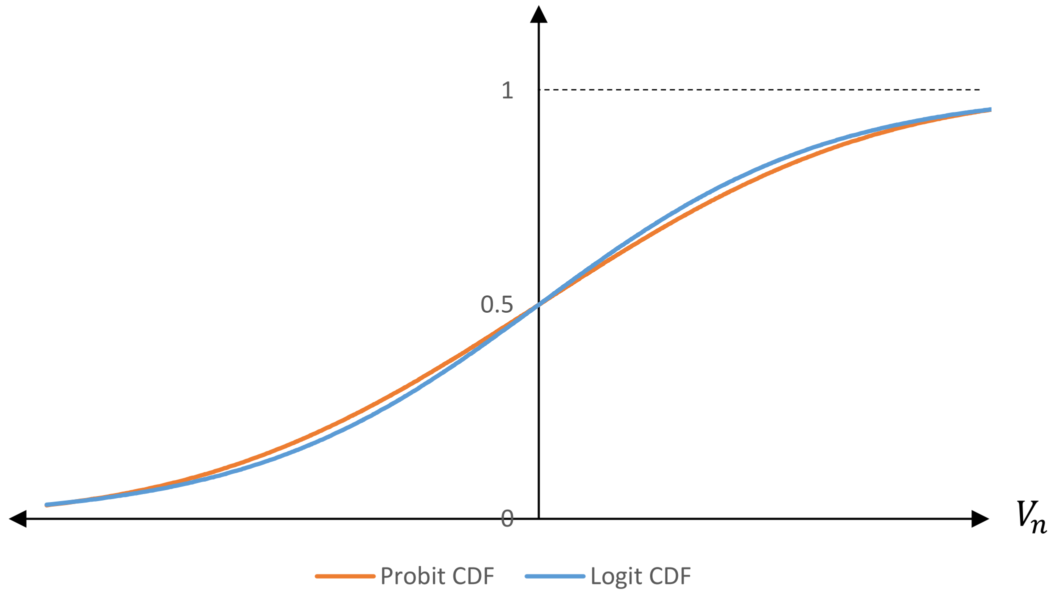
Probabilities are never zero or one.

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

Discrete choice theory

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

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

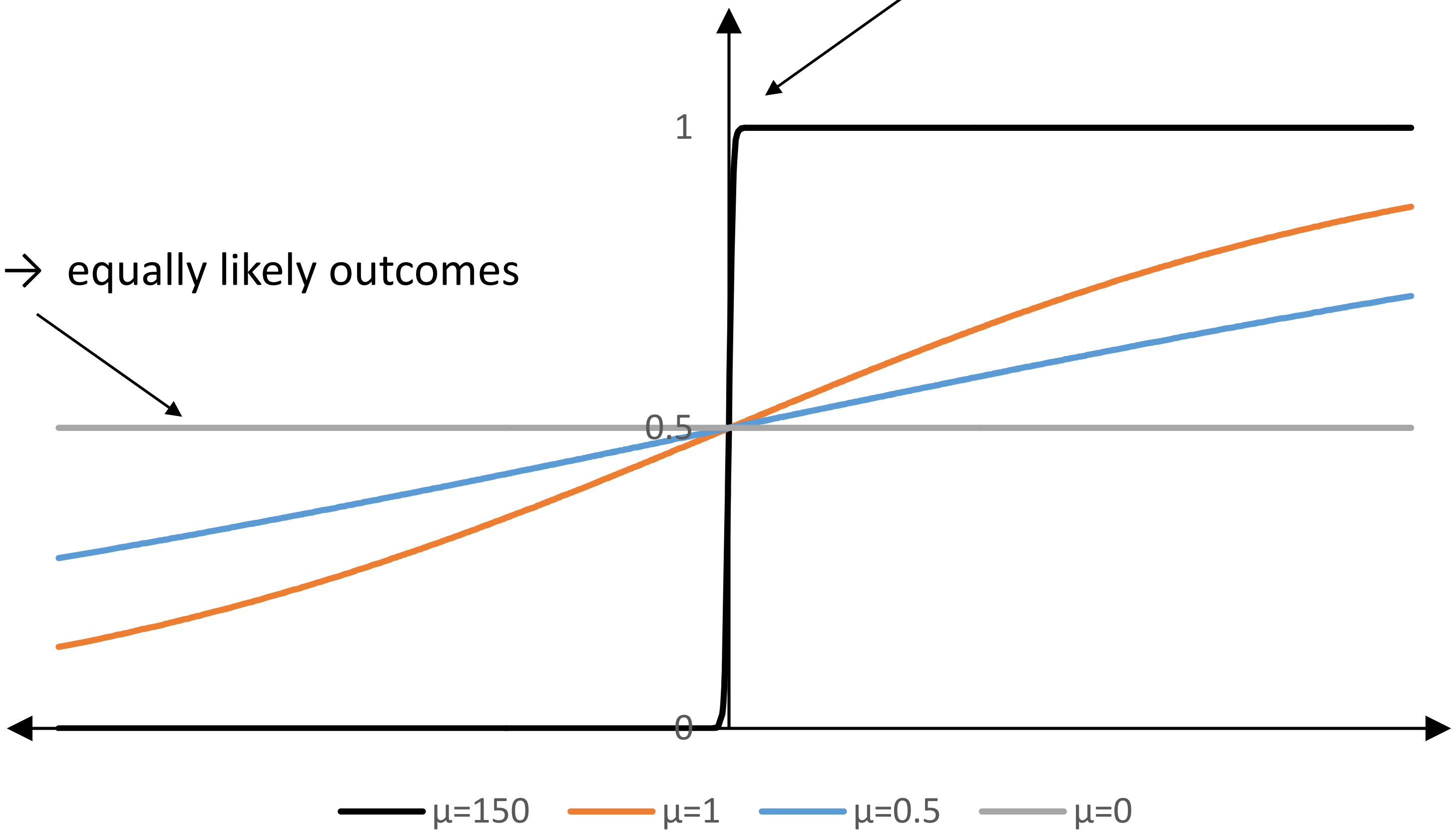
where μ is a scale parameter

Logit Models

An intuitive way of thinking about the scale parameter

As μ approaches infinity → deterministic outcomes

As μ approaches 0 → equally likely outcomes



Where μ is inversely proportional to the variance of the error term.

Discrete choice theory

A mode choice example

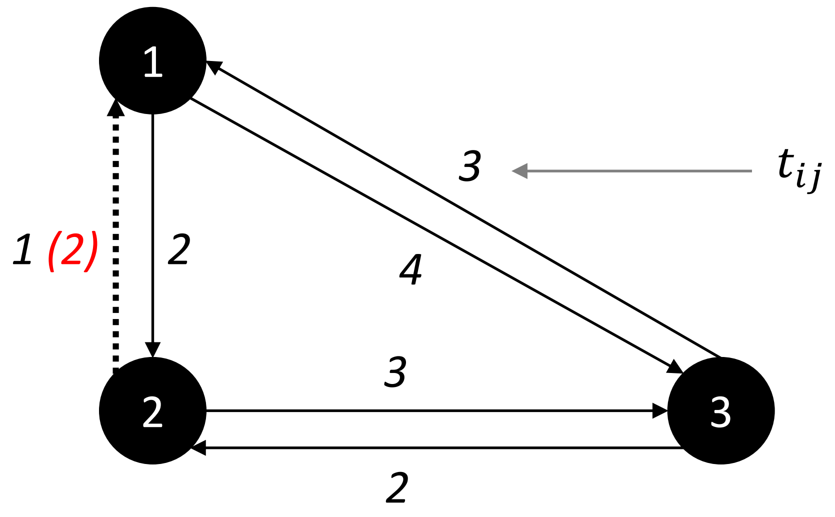
- A binary logit model application *(we will go into more detail later on)*

$$P(Car) = \frac{e^{V_{car}}}{e^{V_{car}} + e^{V_{train}}}$$

$$V_{car} = \beta_{car} + \beta_{costc} Cost_{car} \quad = \quad 1.45 - 0.03\ Cost_{car}$$

$$V_{train} = \beta_{costt} Cost_{train} \quad = \quad -0.01\ Cost_{train}$$

Where V_{car} and V_{train} are utility functions



Value in parenthesis is train cost

Consider the mode choice from zone 2 to zone 1

$$P(Car) = \frac{e^{V_{car}}}{e^{V_{car}} + e^{V_{train}}} = \frac{e^{1.45 - 0.03 \cdot 1}}{e^{1.45 - 0.03 \cdot 1} + e^{-0.01 \cdot 2}}$$

$$\boldsymbol{P(Car) = 81\%} \qquad \boldsymbol{P(Train) = 1 - P(Car)}$$

Logit Models

Regarding the deterministic component of the utility function

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

Only difference in utility matters!

- **Types of variables that go into V:**
- Consider the following utility functions of a binary logit model

Alternative specific constants (ASC's)

- With J alternatives, can only include J-1 constants, one must be normalized to 0
- Reflects the average effect of factors not included in V in relation to the normalized constant.

Alternative specific variables

$$V_{car} = ASC_{car} + \beta_{time} TravelTime + IVC\beta_{csinc}\left(\frac{Cost}{Income}\right) + \gamma_{worker} Worker$$

$$V_{train} = 0 + \beta_{time} TravelTime + IVC\beta_{csinc}\left(\frac{Cost}{Income}\right) + 0$$

- If entered independently, one parameter must be normalized to 0 (similar to ASCs)
- If interacted with Alternative specific variables, no normalization required

Individual specific variables (socio-demographics)

Logit Models

Another example of model choice: 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

LL(0): Log-Likelihood when all parameters are 0

LL(β): Maximum Log-Likelihood

-2[LL(0)-LL(β)]: Test of null hypothesis that all parameters are jointly zero. χ^2 distributed

ρ^2: Informal goodness-of-fit measure : $1 - (LL(\beta)/LL(0))$

$\bar{\rho}^2$: Informal goodness-of-fit measure: $1 - (LL(\beta)-K)/LL(0)$

Adapted from Ben-Akiva and Lerman (1984)

Logit Models

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 derive the MNL

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

$$P(i) = Pr[V_{in} + \varepsilon_{in} \geq V_n^* + \varepsilon_n^*] \longleftarrow$$

$(V_n^* + \varepsilon_n^*)$ is gumbel distributed with parameters $\left(\frac{1}{\mu} ln \sum_{j=1}^{J} \exp(\mu V_{jn}), \mu\right)$

$$P(i) = Pr\left[(V_{jn}^* + \varepsilon_{jn}^*) - (V_{in} + \varepsilon_{in}) \leq 0\right] \longleftarrow$$

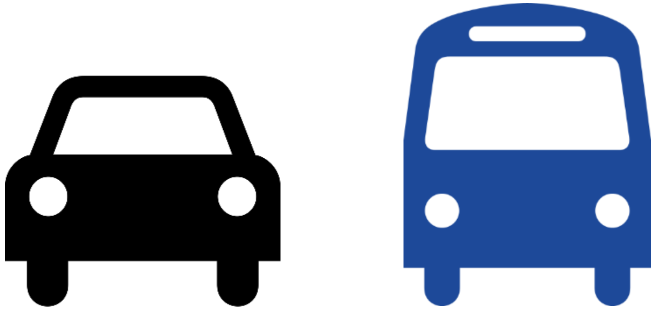
The difference between two Gumbel-distributed variables is Logistic-distributed

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

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

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

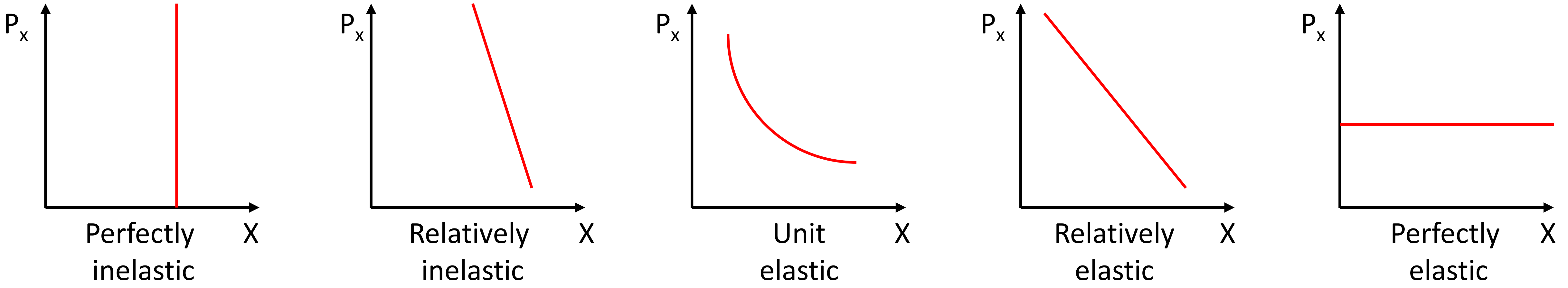
Sample cross-elasticity

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 Elasticities (Point elasticities)

Relation between elasticity of demand, change in price and revenue



	Perfectly inelastic	Relatively inelastic	Unit elastic	Relatively elastic	Perfectly elastic
Direct elasticity:	1% increase in X results in a 0% decrease in P_i	1% increase in X results in a less than 1% decrease in P_i	1% increase in X results in a 1% decrease in P_i	1% increase in X results in a more than 1% decrease in P_i	1% increase in X results in a ∞ decrease in P_i
Cross elasticity:	1% increase in X results in a 0% increase in P_j	1% increase in X results in a less than 1% increase in P_j	1% increase in X results in no percent change in P_j	1% increase in X results in a more than 1% increase in P_j	1% increase in X results in a ∞ increase in P_j

Adapted from Hensher, David A., John M. Rose, and William H. Greene. *Applied choice analysis: a primer*. Cambridge University Press, 2015 (2nd Edition)

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.

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

- **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.

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

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

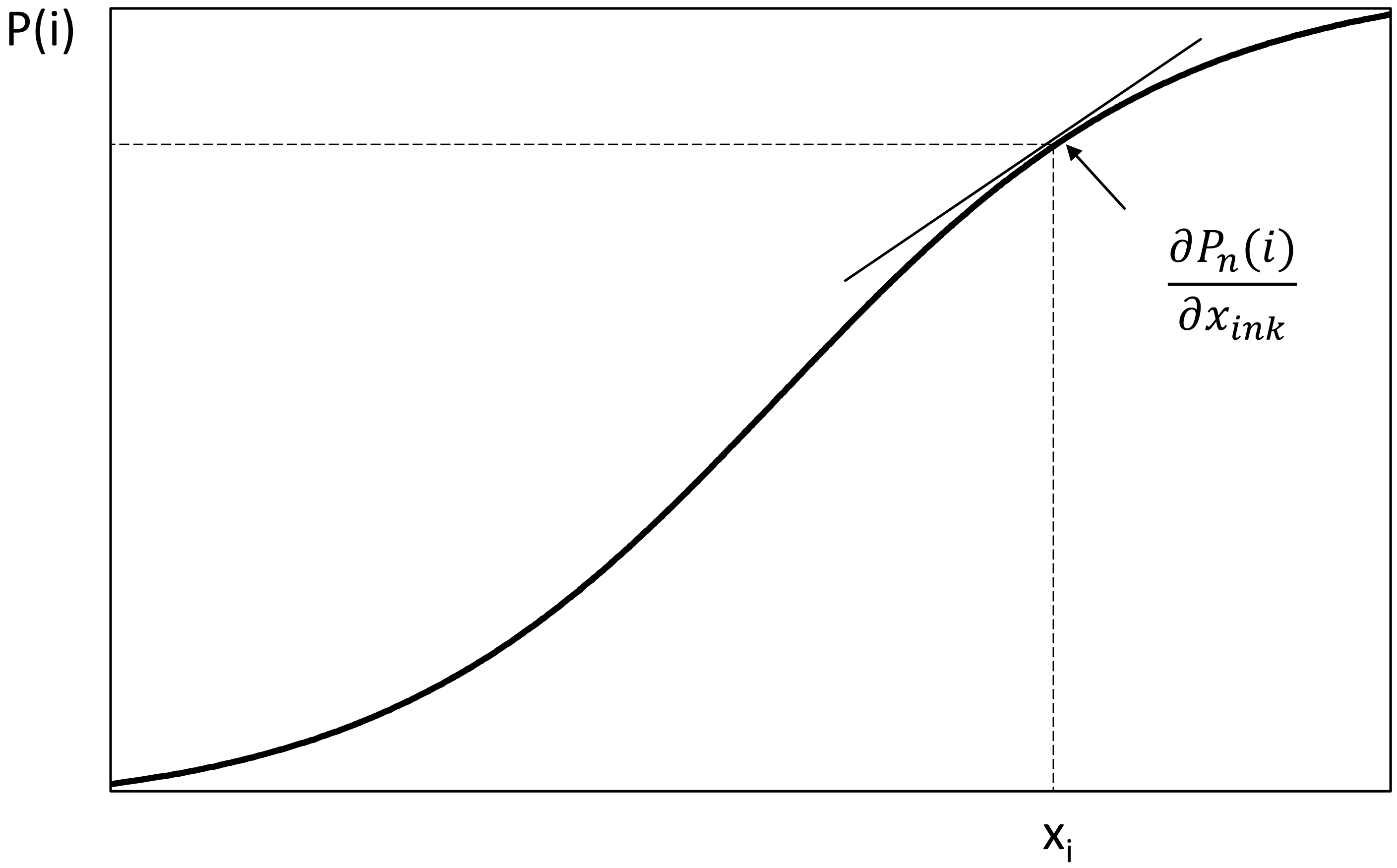
$$M_{x_{jnk}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) M_{x_{jnk}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

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!

MNL: Logit Marginal Effects



Marginal effects as the slopes of the Tangent lines to the cumulative probability curve

Adapted from Hensher, David A., John M. Rose, and William H. Greene. *Applied choice analysis: a primer*. Cambridge University Press, 2015 (2nd Edition)

Strengths and limitations of logit models

The logit model can represent systematic taste variation (related to the observed characteristics of the decision maker),but not random taste variations (linked to unobserved characteristics)

Due to the IIA constraint, logit models can only handle proportional substitution across alternatives, given the researcher’s specification of the utility function. More flexible forms require different models.

The logit model can capture the dynamics of repeated choices if unobserved factors are independent over time only.

Adapted from Train (2002)

Logit Models: Estimation

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:

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

The likelihood is proportional the product of individual probabilities

Maximization of the Log-likelihood function

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

Maximum likelihood estimation of parameters

In the general case, the likelihood function can be defined as the probability that individual n chooses the alternative he was observed choosing.

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

y_{in} takes value 1 when alternative i is chosen, 0 otherwise

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

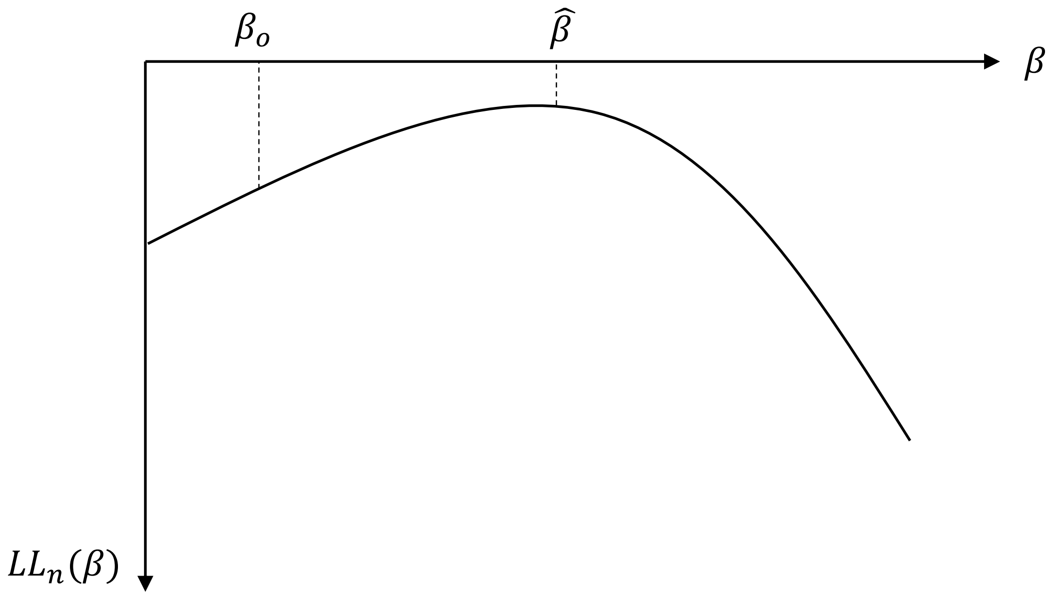
$$LL_n(\beta_1, \beta_2, \dots \beta_K) = \sum_{n=1}^{N} \sum_{i} y_{in} log P_n(i)$$

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)

$$\frac{\partial LL}{\partial \widehat{\beta_k}} = 0, for\ k = 1, \ldots, K$$

At the maximum likelihood, its derivative with respect to each parameter is 0.



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 $\boldsymbol{\beta}$.

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

- All its eigenvalues are non-positive or,
- Its leading principal minors are positive

*In the case of a single variable, this is equivalent to the second derivative test. $f'(c) = 0, f''(x) \leq 0$

Logit Models : Estimation

Maximum likelihood estimation of parameters (Logit Case)

The Log-likelihood function is

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

$$= \sum_{n=1}^{N} \sum_{i} y_{in} \log\left(\frac{\exp(\beta x_{in})}{\sum_{j \in C} \exp(\beta x_{jn})}\right)$$

$$= \sum_{n=1}^{N} \sum_{i} y_{in} \beta x_{in} - \sum_{n=1}^{N} \sum_{i} y_{in} \log\left[\sum_{j \in C} \exp(\beta x_{jn})\right]$$

Logit Models : Estimation

Maximum likelihood estimation of parameters

The FOC is defined as

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

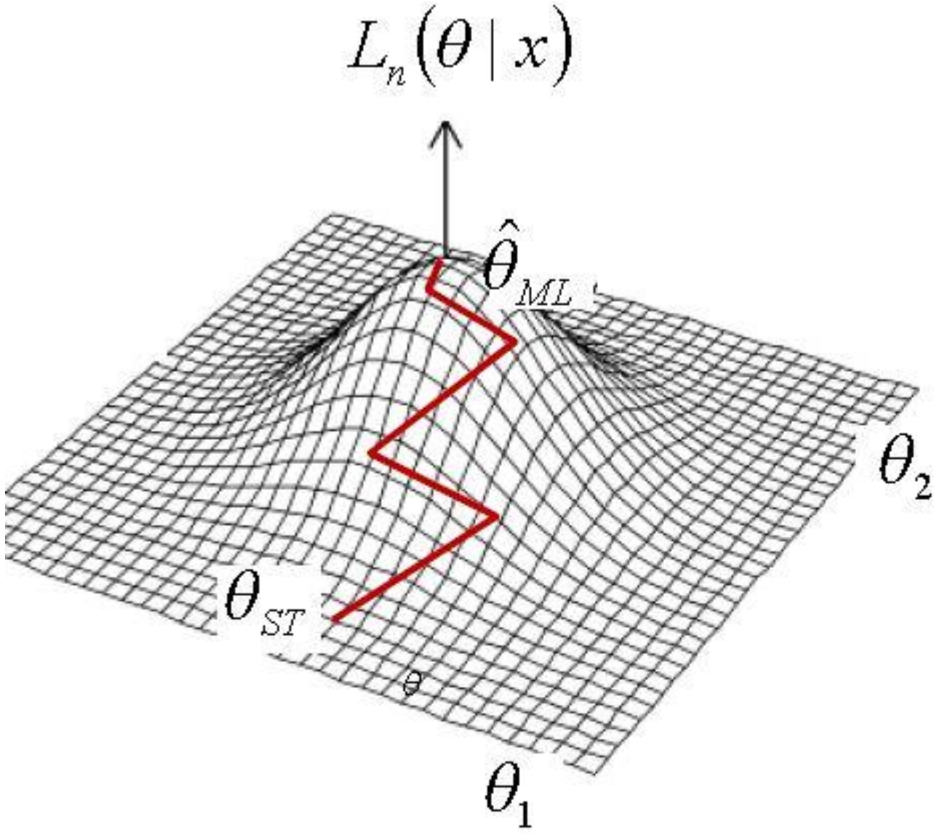
While the second derivatives can be solved as

$$\frac{\partial^2 LL}{\partial \widehat{\beta_k} \partial \widehat{\beta_l}} = -\sum_{n=1}^{N} \sum_{i} P(i) \left[x_{ink} - \sum_{j} x_{jnk} P_n(j) \right] \left[x_{inl} - \sum_{j} x_{jnl} P_n(j) \right] \qquad for\ k = 1, \dots, K$$

Logit Models : Estimation

Maximum likelihood estimation of parameters

- 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) Algorithm



Practical issues in discrete choice modeling

THE UNIVERSITY OF TOKYO

Practical issues in discrete choice modeling

Part I: 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*
    - $P(i|\mathbf{x}_n)$: *the probability of individual n choosing alternative i given attributes* $\mathbf{x}_n$

Practical issues in discrete choice modeling

Part I: Aggregate forecasting techniques

Sample attributes in a MNL mode choice model

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 usually unknown, and even when known, evaluating this integral might be computationally burdensome.

Part I: 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 on the **sample enumeration method** as it is the most widely used.

Practical issues in discrete choice modeling

Part I: Aggregate forecasting techniques

① 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)**

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

Practical issues in discrete choice modeling

Part I: Aggregate forecasting techniques

① 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.

Practical issues in discrete choice modeling

Part II: 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.

Practical issues in discrete choice modeling

Part II: 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

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

Practical issues in discrete choice modeling

Part II: Relevant statistical tests

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

- $\mathrm{H_o}\colon\ \beta_1 = \beta_2 = \cdots = \beta_K = 0$ ←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: $-2\left(LL(\boldsymbol{C}) - LL(\widehat{\boldsymbol{\beta}})\right)$

    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: $-2\left(LL(\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}}) - LL(\widehat{\boldsymbol{\beta}}_{\boldsymbol{u}})\right)$

    Where $LL(\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}})$ is the Log-likelihood of the restricted model, $LL(\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}})$ 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.

Part II: Relevant statistical tests

③ Goodness of fit test:

$$\rho^2 = 1 - \frac{LL(\widehat{\boldsymbol{\beta}})}{LL(0)}$$

← Used in a similar manner to R2 in OLS regression.

$$\bar{\rho}^2 = 1 - \frac{LL(\widehat{\boldsymbol{\beta}}) - K}{LL(0)}$$

← 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.

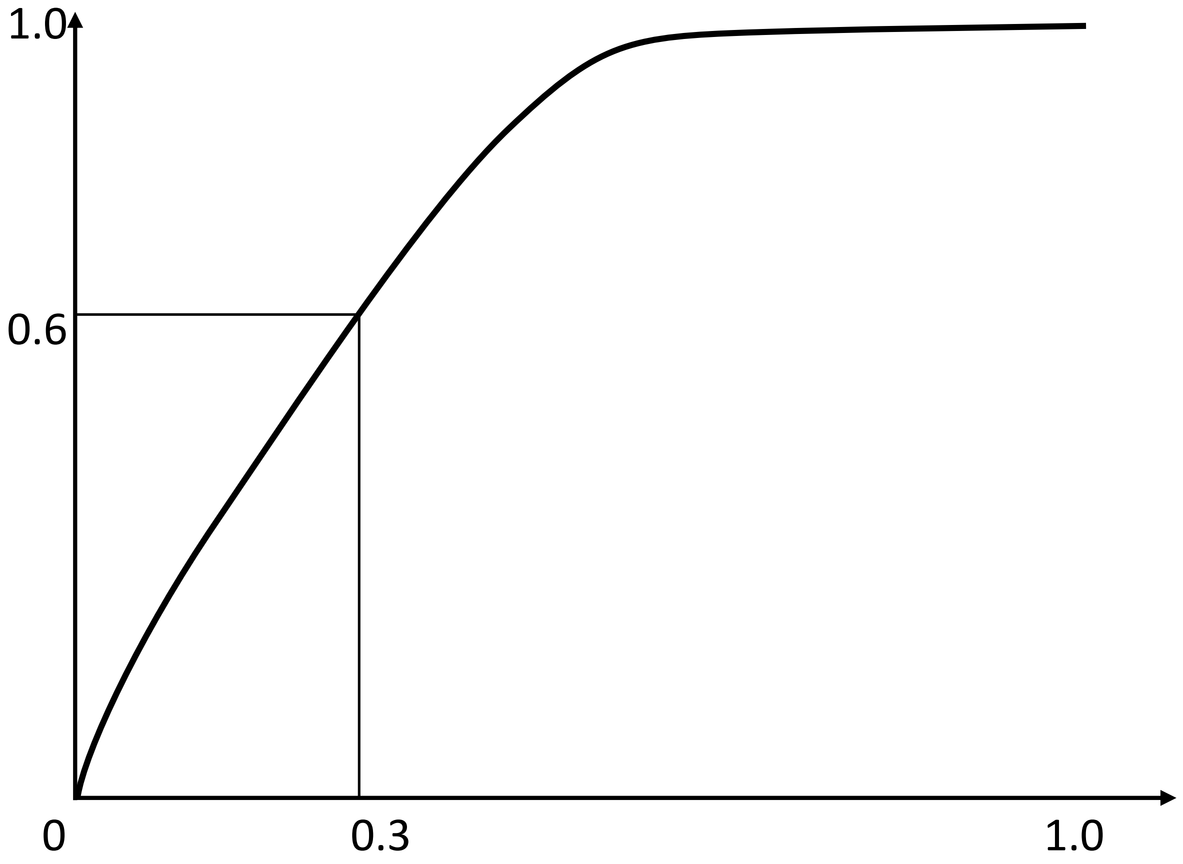
Practical issues in discrete choice modeling

Part II: Relevant statistical tests

③ Goodness of fit test:

- Hensher, Rose and Greene (2015) suggest that a ρ^2 of 0.3 represents a decent model fit for a discrete choice model (approximately 0.6 for R^2 in OLS models).

- ρ^2 ranging from 0.3~0.4 can be translated to R^2 values of 0.6~0.8.



Relationship between R^2 and ρ^2
(Adapted from Hensher, Rose and Greene (2015))

Practical issues in discrete choice modeling

Part II: 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

Part II: 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 $\mathrm{H_o}$: $\boldsymbol{\beta}^{\mathbf{1}} = \boldsymbol{\beta}^{\mathbf{2}} = \cdots = \boldsymbol{\beta}^{\mathbf{G}}$

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

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 gth data subset

Practical issues in discrete choice modeling

Part II: 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 $\quad X^2_{0.05} = 21.0$

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

Individual coefficients can also be compared across Segments:

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

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 (-5.846)	-3.240 (-2.436)
Shared ride (SR) constant	-1.140 (-3.826)	-2.980 (-2.463)
Round-trip travel time (min)	0.028 (3.500)	-0.049 (-2.455)
Round-trip out-of-vehicle time (min)/ one-way distance (0.01 mile)	-14.700 (-2.341)	-14.500 (-1.295)
Cars/workers in household (DA specific)	-35.300 (-1.929)	-35.400 (1.009)
Cars/workers in household (SR specific)	4.260 (9.861)	3.560 (3.849)
Downtown workplace dummy (DA specific)	1.400 (4.106)	2.590 (2.776)
Downtown workplace dummy (SR specific)	-0.605 (-1.644)	-1.130 (-1.865)
Disposable household income (DA specific)	-0.446 (-1.502)	-0.636 (-1.102)
Disposable household income (SR specific)	0.000 (1.335)	0.001 (24.901)
Government worker dummy (SR specific)	0.687 (3.435)	0.063 (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)

Thank you

Questions?

gtroncoso@ut.t.u-Tokyo.a.jp

THE UNIVERSITY OF TOKYO