Discrete Choice Models

CIVL 7012/8012

Discrete Choice Introduction *(1)*

- Discrete or nominal scale data often play a dominant role in transportation
  - because many interesting analyses deal with such data.
- Examples of discrete data in transportation include
  - the mode of travel (automobile, bus, rail transit),
  - place to relocate (urban, sub-urban, local)
  - lane changing (lane to left, right or stay on the same lane)
  - the type or class of vehicle owned, and
  - the type of a vehicular crash (run-off-road, rear-end, head-on, etc.).

Discrete Choice Introduction *(2)*

- From a conceptual perspective,
  - such data are classified as those involving a behavioral choice (choice of mode or type of vehicle to own) or
  - those simply describing discrete outcomes of a physical event (type of vehicle accident).

Models for Discrete Data

- The concept of discrete choice model is
  - the individual decision maker who, faced with a set of feasible discrete alternatives, selects the one that yields greatest utility
  - A set of discrete alternatives form a choice set
- For a variety of reasons the utility of any alternative is, from the perspective of the analyst, best viewed as a random variable.

Random Utility

- In a random utility model the probability of any alternative *i* being selected by person *n* from choice set *Cn* is given by

$$P(i|\mathcal{C}_n) = \Pr(U_{in} \geq U_{jn}, \forall j \in \mathcal{C}_n).$$

- Where
  - *i*, and *j* are two alternatives
  - *Uin*->utility of alternative *i* as perceived by decision maker n
  - *Cn*-> choice set

Random Utility

- We ignore situations where $U_{in} = U_{jn}$ for any i and j in the choice set because
  - if U_{in} and U_{jn} are continuous random variables then the probability $Pr(U_{in} = U_{jn})$ that they are equal is zero.

- Let us pursue the basic idea further by considering the special case where the choice set C_n contains exactly two alternatives.
  - Such situations lead to what are termed binary choice models.

Random Utility

- For convenience we denote the choice set *Cn* as *{i, j}*, where, for example,
  - alternative *i* might be the option of driving to work and
  - alternative *j* would be taking the train.
- The probability of person n choosing *i* is

$$P_n(i) = \Pr(U_{in} \geq U_{jn}),$$

- the probability of choosing alternative j is

$$P_n(j) = 1 - P_n(i).$$

Binary Choice

- Let us develop the basic theory of random utility models into a class of operational binary choice models
- A detailed discussion of binary models serves a number of purposes.
  - First, the simplicity of binary choice situations makes it possible to develop a range of practical models, which is more tedious in more complicated choice situations.
  - Second, there are many basic conceptual problems that are easiest to illustrate in the context of binary choice.
  - Many of the solutions can be directly applied to situations with more than two alternatives.

Systematic component and disturbances

- Uin and Ujn are random variables, we begin by dividing each of the utilities into two additive parts as follows

$$
\begin{aligned}
U_{in} &= V_{in} + \varepsilon_{in}, \\
U_{jn} &= V_{jn} + \varepsilon_{jn}.
\end{aligned}
$$

- Where
  - Vin and Vjn are called the systematic (or representative) components of the utility of i and j;
  - εin and εjn are the random parts and are called the disturbances (or random components).

Specification of the Systematic Component

- If we denote $\beta^T = (\beta 1, \beta 2, \ldots, \beta K)$ as the (row) vector of K unknown

$$
\begin{aligned}
V_{in}(x_{in}, \beta) &= \beta^T x_{in} = \beta_1 x_{in1} + \beta_2 x_{in2} + \cdots + \beta_K x_{inK}, \\
V_{jn}(x_{jn}, \beta) &= \beta^T x_{jn} = \beta_1 x_{jn1} + \beta_2 x_{jn2} + \cdots + \beta_K x_{jnK}.
\end{aligned}
$$

- When such a linear formulation is adopted, parameters $\beta 1, \ldots, \beta K$ are called coefficients.

Specification of the Systematic Component

- A coefficient appearing in all utility functions is generic,
- And a coefficient appearing in only one utility function is alternative specific.
- Consider a binary mode choice example, where one alternative is auto (A) and the other is transit (T), and where the utility functions are defined as

$$\begin{aligned} V_{An} &= 0.37 - 2.13t_{An} \\ V_{Tn} &= \quad\;\; - 2.13t_{Tn}. \end{aligned}$$

Specification of the Systematic Component

- In this case it appears as though the auto utility has an additional term equal to 0.37. We can "convert" this model into the form of equation by defining our x's as follows

$$\begin{aligned} x_{An1} &= 1, \\ x_{Tn1} &= 0, \\ x_{An2} &= t_{An}, \\ x_{Tn2} &= t_{Tn}, \end{aligned}$$

- with K = 2, β1 = 0.37 is alternative specific, and β2 = −2.13 is generic. Thus

$$\begin{aligned} V_{An} &= \beta^T x_{An} &= \beta_1 x_{An1} + \beta_2 x_{An2} &= 0.37 - 2.13 t_{An}, \\ V_{Tn} &= \beta^T x_{Tn} &= \beta_1 x_{Tn1} + \beta_2 x_{Tn2} &= \quad\; - 2.13 t_{Tn}. \end{aligned}$$

Specification of the Systematic Component

- In this example, the variable xAn1 is an alternative specific (i.e., auto) dummy variable and ß1 is called an alternative specific constant.

Linearity in Parameters

- A model with a linear-in-parameter formulation can be described in a specification table.
- A specification table has
  - as many columns as alternatives in the model (two in the specific context of binary choice), and
  - as many rows as coefficients (K).
  - Entry (k, i) of the table contains xik, the variable k for alternative i.

		Auto	Train
β_1	0.37	1	0
β_2	-2.13	t_{An}	t_{Tn}

Linearity in Parameters

- Linearity in the parameters is not as restrictive an assumption as one might first think. Linearity in the parameters is not equivalent to linearity in the variables z and S.
- We allow for any function h of the variables so that polynomial, piecewise linear, logarithmic, exponential, and other transformations of the attributes are valid for inclusion as elements of x.

Illustrative Example

- Let us consider the same example of choosing between auto and transit

$$\begin{aligned} U_{An} &= \beta_0 + \beta_1 t_{An} + \beta_2 c_{An}, \\ U_{Tn} &= \beta_1 t_{Tn} + \beta_2 c_{Tn}, \end{aligned}$$

- Let us consider the traveler has only information about time and not the cost.
- So the cost is added to the error term.
- Depending on what unobserved variables we have the distribution of the error term will change.
- Let us explore more on the functional forms later.

Common Binary Choice Models

- Let us derive operational models by introducing
- the most common binary choice models:
  - the binary probit and
  - the binary logit models.

- In each subsection we begin by making some assumption about the distribution of the two disturbances, εin and εjn, or about the difference between them.
- Given one of these assumptions, we then solve for the probability that alternative i is chosen.

Common Binary Choice Models

- Let us respecify the random utility model

$$\begin{aligned} P_n(i) &= \Pr(\varepsilon_{jn} - \varepsilon_{in} \le V_{in} - V_{jn}) \\ &= \Pr(\varepsilon_n \le V_{in} - V_{jn}), \end{aligned}$$

- Where $\varepsilon_n = \varepsilon_{in} - \varepsilon_{jn}$

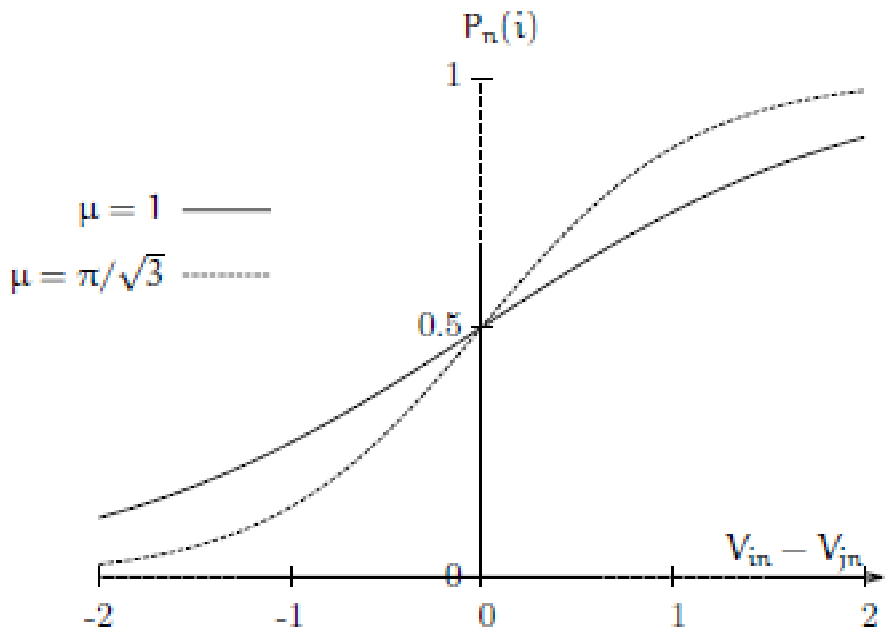
- It means that the probability for individual n to choose alternative i is equal to the probability that the difference Vin – Vjn exceeds the value of εn.
- We need to know how εn is distributed

Binary Logit

- For binary logit the choice probability for alternative *i* is given by

$$
\begin{aligned}
P_n(i) &= \Pr(\varepsilon_n \le V_{in} - V_{jn}) \\
&= F(V_{in} - V_{jn}) \\
&= \frac{1}{1 + e^{-\mu(V_{in} - V_{jn})}} \\
&= \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}}.
\end{aligned}
$$

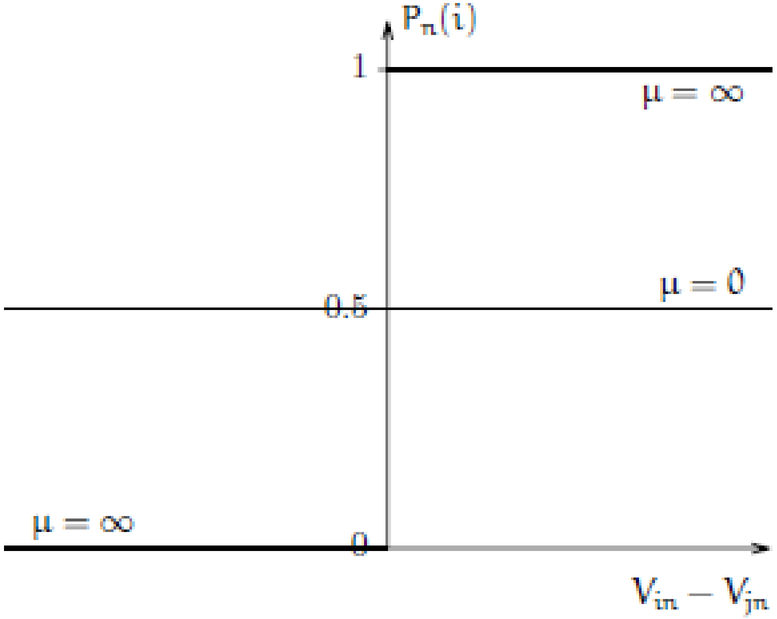
Binary Logit Shape



Limiting Case of Binary Logit

- If Vin and Vjn are linear in their parameters
- μ is the scale parameter

$$P_n(i) = \frac{e^{\mu\beta^T x_{in}}}{e^{\mu\beta^T x_{in}} + e^{\mu\beta^T x_{jn}}} = \frac{1}{1 + e^{-\mu\beta^T (x_{in} - x_{jn})}}.$$



Limiting Case of Binary Logit

- In the case of linear-in-parameters utilities, the parameter μ cannot be distinguished from the overall scale of the β's.
- For convenience we generally make an arbitrary assumption that $\mu = 1$.
- This corresponds to assuming the variances of εin and εjn are both $\pi^2/6$, implying that $Var(\varepsilon jn - \varepsilon in) = \pi^2/3$.

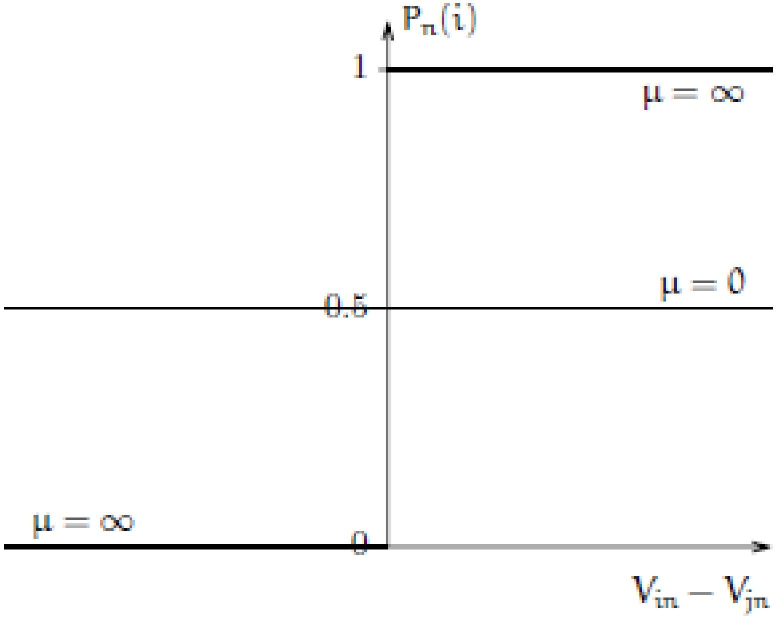
Limiting Case of Binary Logit

- Note that this differs from the standard scaling of binary probit models, where we set $Var(\varepsilon_{jn} - \varepsilon_{in}) = 1$, and it implies that the scaled logit coefficients are $\pi/\sqrt{3}$ times larger than the scaled probit coefficients.
- A rescaling of either the logit or probit utilities is therefore required when comparing coefficients from the two models.

Limiting Case of Binary Logit

- that is, as μ → ∞, the choice model is deterministic. On the other hand,
- when μ → 0, the choice probability of i becomes 1/2

$$\lim_{\mu \to \infty} P_n(i) = \begin{cases} 1 & \text{if } V_{in} - V_{jn} > 0, \\ 0 & \text{if } V_{in} - V_{jn} < 0; \end{cases}$$



Estimation Approach

- The model coefficients reflect the sensitivity of the behavior to the variables.
- To identify them, we use data on behavioral choices describing individuals, what they faced, and what they chose.
- Therefore, we turn now to the problem of estimating the values of the unknown parameters ß1,. . . ,ßK from a sample of observations.

Estimation Approach

- Each observation consists of the following

An indicator variable defined as

$$y_{in} = \begin{cases} 1 & \text{if person } n \text{ chose alternative } i, \\ 0 & \text{if person } n \text{ chose alternative } j. \end{cases}$$

- Two vectors of attributes xin = h(zin, Sn) and xjn = h(zjn, Sn), each containing K values of the relevant variables.

Estimation Approach

- Given a sample of N observations, our problem then becomes one of finding estimates $\hat{\beta}_1, \ldots, \hat{\beta}_K$ that have some or all of the desirable properties of statistical estimators.
- We consider in detail the most widely used estimation procedure — maximum likelihood.

Maximum Likelihood

- The maximum likelihood estimation (MLE) procedure is conceptually quite straightforward.
- It consists in identifying the value of the unknown parameters such that the joint probability of the observed choices as predicted by the model is the highest possible.
- This joint probability is called the likelihood of the sample.

Maximum Likelihood

- Consider the likelihood of a sample of N observations assumed to be independently drawn from the population.
- The likelihood of the sample is the product of the likelihoods (or probabilities) of the individual observations
- Let us define the likelihood function as

$$\mathcal{L}^*(\beta_1, \beta_2, \ldots, \beta_K) = \prod_{n=1}^{N} P_n(i)^{y_{in}} P_n(j)^{y_{jn}},$$

- Where, Pn(i) and Pn(j) are functions of ß1,. . . ,ßK.

Maximum Likelihood

- Note

$$P_n(i)^{y_{in}} P_n(j)^{y_{jn}} = \begin{cases} P_n(i) & \text{if } y_{in} = 1, y_{jn} = 0 \\ P_n(j) & \text{if } y_{in} = 0, y_{jn} = 1. \end{cases}$$

- The log likelihood is written as follows

$$\mathcal{L}(\beta_1, \ldots, \beta_K) = \sum_{n=1}^{N} (y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)),$$

- Noting that

noting that $y_{jn} = 1 - y_{in}$ and $P_n(j) = 1 - P_n(i)$,

Maximum Likelihood

- The log-likelihood function is given by

$$\mathcal{L}(\beta) = \mathcal{L}(\beta_1, \ldots, \beta_K) = \sum_{n=1}^{N} (y_{in} \ln P_n(i) + (1 - y_{in}) \ln(1 - P_n(i))),$$

- Maximize the log-likelihood

$$\max \mathcal{L}(\hat{\beta}) = \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K),$$

- First order conditions

$$\frac{\partial \mathcal{L}}{\partial \beta_k}(\hat{\beta}) = \sum_{n=1}^{N} \left(y_{in} \frac{\partial P_n(i)/\partial \beta_k}{P_n(i)} + y_{jn} \frac{\partial P_n(j)/\partial \beta_k}{P_n(j)} \right) = 0, \; k = 1, \ldots, K,$$

$$\frac{\partial \mathcal{L}}{\partial \beta}(\hat{\beta}) = 0.$$

Maximum Likelihood

- Each entry k of the vector $\partial L(b\beta)/\partial \beta$ represents the slope of the multi-dimensional log likelihood function along the corresponding kth axis.
- If $b\beta$ corresponds to a maximum of the function, all these slopes must be zero
- Essentially an optimization problem requires efficient techniques to solve for estimates

Example-1: Netherland Mode Choice

- The example deals with mode choice behavior for intercity travelers in the city of Nijmegen (the Netherlands) using revealed preference data.
- The survey was conducted during 1987 for the Netherlands Railways to assess factors that influence the choice between rail and car for intercity travel

Example-1: Netherland Mode Choice

	Car	Train
β_1	1	0
β_2	cost of trip by car (in Guilders)	cost of trip by train (in Guilders)
β_3	travel time by car (hours) if trip purpose is work, 0 otherwise	0
β_4	travel time by car (hours) if trip purpose is not work, 0 otherwise	0
β_5	0	travel time by train (hours)
β_6	0	1 if first class is preferred, 0 otherwise
β_7	1 if commuter is male, 0 otherwise	0
β_8	1 if commuter is the main earner in the family, 0 otherwise	0
β_9	1 if commuter had a fixed arrival time, 0 otherwise	0

Example-1: Netherland Mode Choice

- Coefficient ß1 is the alternative specific constant
- ß2 is the coefficient of travel cost
- ß3 and ß4 are coefficients of car travel time.
- ß5 is the coefficient of train travel time
- Coefficient ß6 measures the impact on the utility of the train if the class preference for rail travel is first class.
- ß7, ß8 and ß9 are coefficients of alternative-

Example-1: Netherland Mode Choice

- Input data format

	Individual 1	Individual 2	Individual 3
Train cost	40.00	7.80	40.00
Car cost	5.00	8.33	3.20
Train travel time	2.50	1.75	2.67
Car travel time	1.17	2.00	2.55
Gender	M	F	F
Trip purpose	Not work	Work	Not work
Class	Second	First	Second
Main earner	No	Yes	Yes
Arrival time	Variable	Fixed	Variable

Binary Probit

			Individual 1		Individual 2		Individual 3	
Variables	Coef.	Value	Car	Train	Car	Train	Car	Train
Car dummy	β_1	1.77	1	0	1	0	1	0
Cost	β_2	-0.0296	5.00	40.00	8.33	7.80	3.20	40.00
Travel time by car (work)	β_3	-1.51	0	0	2.00	0	0	0
Travel time by car (not work)	β_4	-1.26	1.17	0	0	0	2.55	0
Travel time by train	β_5	-0.308	0	2.50	0	1.75	0	2.67
First class dummy	β_6	0.545	0	0	0	1	0	0
Male dummy	β_7	-0.471	1	0	0	0	0	0
Main earner dummy	β_8	0.213	0	0	1	0	1	0
Fixed arrival time dummy	β_9	-0.355	0	0	1	0	0	0
V_{in}			-0.3120	-1.9551	-1.6354	-0.2252	-1.3126	-2.0065
$P_n(i)$			0.950	0.0502	0.0792	0.921	0.756	0.244

$$
\begin{aligned}
P_1(\text{car}) &= \Pr(-0.3120 + \varepsilon_{\text{car}1} \geq -1.9551 + \varepsilon_{\text{train}1}) \\
&= \Pr(1.6431 \geq \varepsilon_1),
\end{aligned}
$$

Binary Probit

- P1(car) = $\Phi(1.6431)$ = 0.950.
- We compute similarly that P2(car) = 0.0792 and P3(car) = 0.756

Binary Logit

			Individual 1		Individual 2		Individual 3	
Variables	Coef.	Value	Car	Train	Car	Train	Car	Train
Car dummy	β_1	3.04	1	0	1	0	1	0
Cost	β_2	-0.0527	5.00	40.00	8.33	7.80	3.20	40.00
Travel time by car (work)	β_3	-2.66	0	0	2	0	0	0
Travel time by car (not work)	β_4	-2.22	1.17	0	0	0	2.55	0
Travel time by train	β_5	-0.576	0	2.50	0	1.75	0	2.67
First class dummy	β_6	0.961	0	0	0	1	0	0
Male dummy	β_7	-0.850	1	0	0	0	0	0
Main earner dummy	β_8	0.383	0	0	1	0	1	0
Fixed arrival time dummy	β_9	-0.624	0	0	1	0	0	0
V_{in}			-0.6642	-3.5504	-2.9596	-0.4589	-2.4072	-3.6464
$P_n(i)$			0.947	0.0528	0.0758	0.924	0.775	0.225

$$P_1(\text{car}) = \frac{e^{-0.6642}}{e^{-0.6642} + e^{-3.5504}} = 0.947,$$

$$P_1(\text{train}) = 1 - P_1(\text{car}) = 0.0528$$

Comparison

- the coefficients of the binary logit must be divided by $\pi/\sqrt{3}$ in order to be compared to the coefficients of the binary probit model

	Logit	Scaled logit	Probit
β_1	3.04	1.68	1.77
β_2	-0.0527	-0.0291	-0.0296
β_3	-2.66	-1.47	-1.51
β_4	-2.22	-1.22	-1.26
β_5	-0.576	-0.318	-0.308
β_6	0.961	0.53	0.545
β_7	-0.85	-0.469	-0.471
β_8	0.383	0.211	0.213
β_9	-0.624	-0.344	-0.355