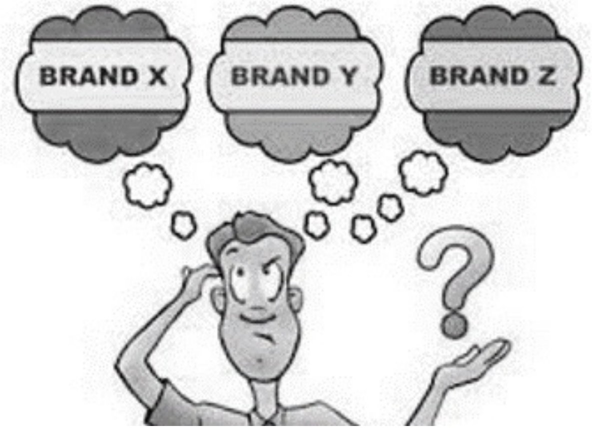


B9651 – Marketing Analytics

Session 6: Choice Models

Professor Hortense Fong

Logistics

- Individual Assignment 1 due Wed, Oct 16 at 8PM
- No classes next week
- Midterm Oct 22 + 23
  - Weeks 1-5, closed-note, calculator allowed

Last Time

- Recommender Systems + Modeling Customer Churn
- Today: Modeling Customer Choices

Today: Modeling Choice

1. How can we model the choice process of customers?
2. What are the different types of choice models?
    1. Binary logit
    2. Multinomial Logit
3. How can we estimate the Logit Models?
4. What are the limitations of Multinomial Logit Models?

Today's Goals

Understand

- What are choice models
- How to construct a utility function
- How to use Maximum Likelihood Estimation to model consumer choices
- The limitations of Multinomial Logit Models

Be able to

- Construct a statistical model of choice
- Estimate a choice model in Excel and Python
- Analyze own- and cross-price elasticities

Course Roadmap

STP Analytics (Identify Value)	Customer Analytics (Deliver Value)	4P Analytics (Capture Value)
Module 1	**Module 2**	**Module 3**
What datasets can we use? How can we segment and target our customers? How should we position our products/services?	How much are our customers worth? Are our customers leaving? How do our customers make choices?	How do we build a new product? How should we price our products? How do we distribute them? How do we quantify the impact of our promotions?

Choice Models

Motivation

Modeling Choice

- Imagine that you work at Megabus
  - Provide regular intercity bus routes at a low cost
- Problem: You don't understand when people choose to take a bus or another transportation mode!
  - Alternatives: Train, Plane, Car

- What do you do?
- What if you want to know how price sensitive people are or how important it is compared to travel time?
- Today: we will build a (choice) model to address this type of question

Choice Models – Other Applications

- Choice Models
  - Describe how people make choices
  - Predict choices under different conditions

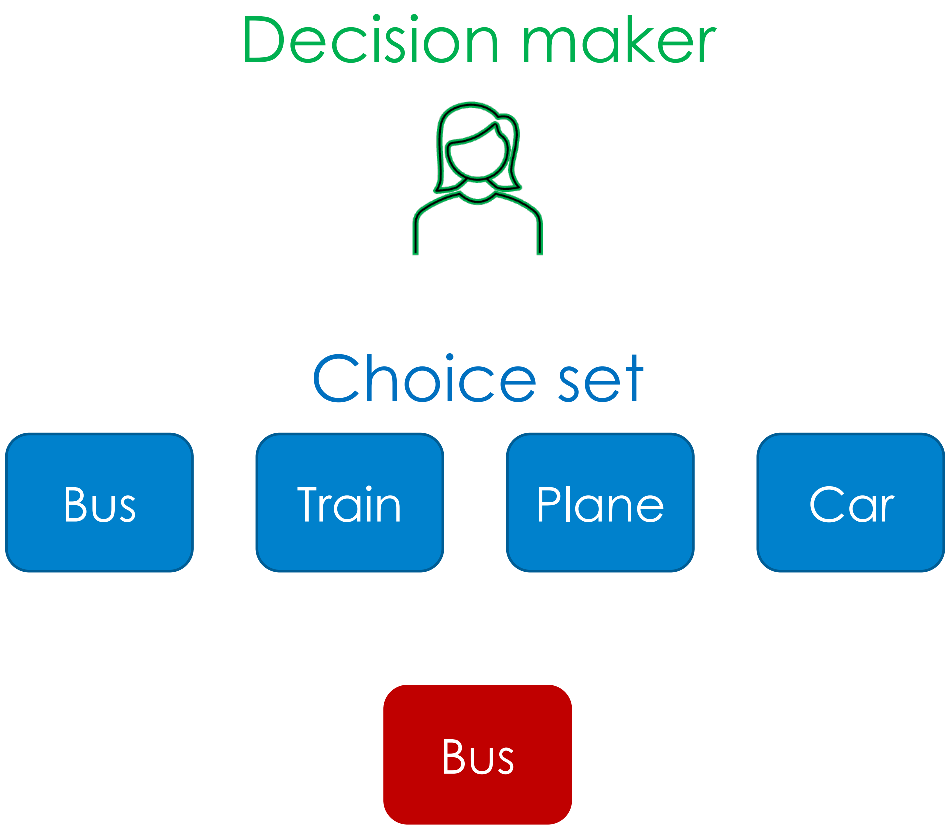
- Common Choice Problems
  - Purchase or Not (Purchase incidence)
  - Brand Choice
  - Credit Default Prediction

Choice Models

Assumptions

Choice Problem

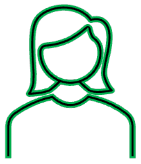
- Choices are made by decision makers
  - Individual Consumers, Households, Firms
- Choice sets contain
  - A finite number of alternatives
  - Alternatives are mutually exclusive
  - Collectively exhaustive
- Choose one alternative from a choice set



Consumers

- Consumer choices are modeled in terms of
  - Characteristics of consumers (age, income, etc.)
  - Attributes of alternatives (price, travel time, etc.)
- Consumers have preferences over the attributes
  - Preferences for attributes are represented by attribute weights
  - Preferences for attributes translate into preferences for alternatives

Decision maker



25 y.o
Female
$80,000

Choice set

Bus	Train	Plane	Car
\$20 5 hours	\$100 4 hours	\$200 1 hour	\$50 4.5 hours

$\beta_1 Price + \beta_2 Time$

Bus

Consumer Decision Rule

- Consumers preferences for alternatives are represented by utility functions
- Utility functions assign one scalar numerical value to each alternative
- Utilities are functions of
  - Attributes of alternatives
  - Characteristics of consumers

  E.g., Transportation choice is function of price, time, and income
- Rational consumers choose the alternative with the highest utility
  - Utility Maximizing Decision Rule

Example: Maximum Utility Choice Rule

- t = time, c = cost, y = income in $K
- Suppose utility is given by $U(t, c, y) = -t - \frac{5c}{y}$

Mode	Time (t) (Hours)	Cost (c) ($)	Utility $y = \$40k$
Drive Alone	0.50	2.00	
Carpool	0.75	1.00	
Bus	1.00	0.75	

Which mode of transportation maximizes utility?

0

Drive Alone

0

Carpool

0

Bus

0

Example: Maximum Utility Choice Rule

- t = time, c = cost, y = income in $K
- Suppose utility is given by $U(t,c,y) = -t - \frac{5c}{y}$

Mode	Time (t) (Hours)	Cost (c) ($)	Utility $y = \$40k$	Utility $y = \$10k$
Drive Alone	0.50	2.00	**-0.75**	-1.50
Carpool	0.75	1.00	-0.88	**-1.25**
Bus	1.00	0.75	-1.09	-1.38

- How can you explain the above utilities for the two travelers?

Effect of Reduction in Bus Travel Time by a Quarter of an Hour

- t = time, c = cost, y = income in $K
- Suppose utility is given by $U(t, c, y) = -t - \frac{5c}{y}$

Mode	Time (t) (Hours)	Cost (c) ($)	Utility $y = \$40k$	Utility $y = \$10k$
Drive Alone	0.50	2.00	**-0.75**	-1.50
Carpool	0.75	1.00	-0.88	-1.25
Bus	0.75	0.75	-0.84	**-1.13**

Choice Models

From Assumptions To Mathematics

Consumer Preferences: Mathematical Representation

- Let $\mathcal{J}$ represent the choice set

- Let $\boldsymbol{w}_{ij}$ represent **all** the attributes of alternative j that consumer i faces
  - E.g., Time, Cost
- Let $\boldsymbol{r}_i$ be the vector of **all** consumer characteristics that are relevant for choice
  - E.g., Income

- Utility is a function $U_{ij} = U(\boldsymbol{w}_{ij}, \boldsymbol{r}_i), \forall j \in \mathcal{J}$

  (for all alternatives in the choice set)
- Consumer decision rule is deterministic:
  - Choose alternative k if $U_{ik} > U_{ij}$ for all $j \neq k, j \in \mathcal{J}$ (Choose the alternative with the highest utility)

The Problem with Deterministic Utility

- The previous travel example used a deterministic utility function
- The function implies consistency in behavior
  - For each consumer, each time the consumer faces the same task
  - For identical consumers
- Real datasets show a lot of inconsistency

Inconsistency in Choice

- What inconsistencies do you see in the example below?

Customer	Age	Price_A	Price_B	Choice
1	23	1.25	1.15	B
1	23	1.25	1.35	A
1	23	1.25	1.15	A
2	25	1.15	1.25	B
2	25	1.25	1.15	A
2	25	1.15	1.25	A
3	31	1	1.15	A
4	23	1.25	1.35	B

Inconsistency in Choice

- Potential sources of inconsistency:
  1. Consumers have incomplete or incorrect information about the attributes
  2. Analyst has incomplete or incorrect information about the attributes or circumstances of the customers

- To account for incomplete information, we use a random utility model to rationalize the observed data

Choice Models

Assumptions (again!) and Random Utility

Random Utility

- Researcher does not know utilities exactly
  - Only few consumer characteristics, z_i out of r_i are known
  - Only few of the attributes, x_{ij} out of w_{ij} are known
- Solution: treat the utility, U_{ij}, as random with additive errors

$$U_{ij} = V_{ij} + \epsilon_{ij}$$

- $V_{ij}(x_{ij}, z_i ; \boldsymbol{\beta})$ is the **systematic** part of the utility
- ϵ_{ij} is the **stochastic** part
  - Represents total impact of all unobserved attributes and demographics relevant to a given choice occasion

Probabilistic Choice

- Given the stochastic part, we can only model choice probabilistically

- Probability of choosing alternative $j \in \mathcal{J}$ by customer i is given by

$$P_{ij} = \text{Prob}\left(U_{ij} > U_{ik}, \forall k \neq j, k \in \mathcal{J}\right)$$

$$P_{ij} = \text{Prob}\left(V_{ij} + \epsilon_{ij} > V_{ik} + \epsilon_{ik}, \forall k \neq j, k \in \mathcal{J}\right)$$

$$P_{ij} = \text{Prob}\left(\underbrace{\epsilon_{ik} - \epsilon_{ij}}_{\text{Random}} < \underbrace{V_{ij} - V_{ik}}_{\text{Systematic}}, \forall k \neq j, k \in \mathcal{J}\right)$$

Example: Two Brands

- Let the utilities be given by
  - $U_{i1} = V_{i1} + \epsilon_{i1}$
  - $U_{i2} = V_{i2} + \epsilon_{i2}$

- Probability of choosing brand 2 for individual i is

$$P_{i2} = \text{Prob}(U_{i1} < U_{i2})$$

$$= \text{Prob}(\epsilon_{i1} - \epsilon_{i2} < V_{i2} - V_{i1})$$

Only differences in utilities matter!

Example: Three Brands

- Let the utilities be given by
  - $U_{i1} = V_{i1} + \epsilon_{i1}$
  - $U_{i2} = V_{i2} + \epsilon_{i2}$
  - $U_{i3} = V_{i3} + \epsilon_{i3}$
- What is the probability of choosing brand 3 in terms of differences in errors?

$$P_{i3} = Prob(U_{i1} < U_{i3}, U_{i2} < U_{i3})$$

$$= Prob(V_{i1} + \epsilon_{i1} < V_{i3} + \epsilon_{i3}, V_{i2} + \epsilon_{i2} < V_{i3} + \epsilon_{i3})$$

$$= Prob(\epsilon_{i1} - \epsilon_{i3} < V_{i3} - V_{i1}, \epsilon_{i2} - \epsilon_{i3} < V_{i3} - V_{i2})$$

Only differences in utilities matter!

Implications

- Only differences in utilities matter

- Implications on the systematic part of the utility
  - What happens when we add the same constant to all utilities?
  - Probabilities don't change: $V_1 = 4, V_2 = 5$ is the same as $V_1 = 2, V_2 = 3$
  - Probability of choosing brand 2 for individual i is

$$
\begin{aligned}
P_{i2} = \text{Prob}(\epsilon_{i1} - \epsilon_{i2} < V_{i2} - V_{i1}) \\
= \text{Prob}(\epsilon_{i1} - \epsilon_{i2} < 5 - 4) \\
= \text{Prob}(\epsilon_{i1} - \epsilon_{i2} < 3 - 2)
\end{aligned}
$$

Choice Models

Systemic Utility

Systematic Part of the Utility

- The systematic part $V_{ij}(.)$ can be written as

$$V_{ij} = V_j(x_{ij}) + V_j(z_i) + V_j(x_{ij}, z_i) + bias_j$$

- consumer i
- alternative j

- $V_j(x_{ij})$ is the portion that is associated with the attributes of alternative j faced by consumer i
- $V_j(z_i)$ is the portion of utility associated with the characteristics of the consumer
- $V_j(x_{ij}, z_i)$ contains the interactions between the attributes and characteristics
- $bias_j$ is the alternative specific constant
- Our entire objective will be to estimate the various components

Portion 1: Systematic Utility Composed of Attributes $\boldsymbol{V_j(x_{ij})}$

- The attribute portion of the systematic utility, $V_j(x_{ij})$ can be written as

$$V_j(x_{ij}) = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_k x_{ijk}$$

- β_k is the coefficient of the kth attribute (e.g., weight of price)
  - **This is what we want to know/estimate!**

- x_{ijk} is the value of the kth attribute (e.g., price faced by consumer i for alternative j)

Portion 1: Systematic Utility Composed of Attributes $V_j(x_{ij})$

- Consider two modes A (airplane) and B (bus) described on two attributes: Price and TravelTime
- The attribute-specific component of the systematic utility is

$$V_A(x_{iA}) = \beta_1 Price_{iA} + \beta_2 TravelTime_{iA}$$

$$V_B(x_{iB}) = \beta_1 Price_{iB} + \beta_2 TravelTime_{iB}$$

- Notice that the coefficients are the same across alternatives
  - Why?
  - Assumption is that Price (and TravelTime) sensitivity is the same across modes

Portion 2: Systematic Component for Individual Characteristics $V_j(z_i)$

- Suppose we have two demographics for consumer i: Income and Family-size
- The individual-specific component can be written as

$$V_A(z_i) = \beta_{1A} Income_i + \beta_{2A} FamilySize_i$$

$$V_B(z_i) = \beta_{1B} Income_i + \beta_{2B} FamilySize_i$$

- Notice that each mode has a different coefficient for each variable.
  - Why? Only differences in utility matter: if it was the same, $V_A(z_i) - V_B(z_i) = \beta_1 Income_i + \beta_2 FamilySize_i - (\beta_1 Income_i + \beta_2 FamilySize_i) = 0$

Portion 3: Systematic Component with Interactions $V_j(x_{ij}, z_i)$

- Different individuals may evaluate attributes differently
- We can interact the attributes with the demographics

$$V_A(x_{iA}, z_i) = \beta_1 Price_{iA} + \beta_2 Price_{iA} \times Income_i$$

$$V_B(x_{iB}, z_i) = \beta_1 Price_{iB} + \beta_2 Price_{iB} \times Income_i$$

- How do we interpret the coefficients?
  - Total effect of a unit increase in price: $\beta_1 + \beta_2 \times Income_i$
  - How much price matters depends on income
    - High income people will be less price sensitive compared to low-income people (when $\beta_1 < 0$ and $\beta_2 > 0$)

Portion 4: Alternative Specific Constants $bias_j$

- The systematic utilities also contain intercepts (or biases) that are alternative specific

$$V_{iA} = \beta_{0A} + \cdots$$
$$V_{iB} = \beta_{0B} + \cdots$$

- These represent the mean of all the unobserved variables ϵ_{ij}
  - Ex. Comfort, safety, privacy (difficult to measure variables)

- They capture the baseline utilities
  - What is unique about each alternative

Putting It All Together: Systematic Utility V_{ij}

- We can assemble all components to get

$$V_{iA} = \beta_{0A} + \beta_{1A} Income_i + \beta_2 Price_{iA} + \beta_3 Price_{iA} \times Income_i$$

$$V_{iB} = \beta_{0B} + \beta_{1B} Income_i + \beta_2 Price_{iB} + \beta_3 Price_{iB} \times Income_i$$

For illustration, we include only one attribute (price) and one consumer characteristic (income)

- Because only differences in utilities matter, not all these coefficients are identifiable (i.e., have a unique value)

unidentifiable

$$\begin{aligned} V_{iA} - V_{iB} = & (\beta_{0A} - \beta_{0B}) \\ & + (\beta_{1A} - \beta_{1B})\, Income_i \\ & + \beta_2 (Price_{iA} - Price_{iB}) \\ & + \beta_3 Income_i \times (Price_{iA} - Price_{iB}) \end{aligned}$$

Many values of the estimates would lead to the same difference in utility

Putting It All Together: Systematic Utility V_{ij}

$$V_{iA} - V_{iB} = (\beta_{0A} - \beta_{0B}) \quad \leftarrow \text{unidentifiable}$$
$$+ (\beta_{1A} - \beta_{1B})\, Income_i$$
$$+ \beta_2 (Price_{iA} - Price_{iB})$$
$$+ \beta_3 Income_i \times (Price_{iA} - Price_{iB})$$

- Because only differences in utility matter, the two intercepts cannot be estimated separately. Only their difference can be estimated.
- Many values of the estimates would lead to the same difference in utility
  - E.g., $\beta_{0A} = 2, \beta_{0B} = 1$ same as $\beta_{0A} = 3, \beta_{0B} = 2$

Putting It All Together: Systematic Utility V_{ij}

- We can set some parameters to zero to obtain identification

$$V_{iA} = 0 + 0 \times Income_i + \beta_2 Price_{iA} + \beta_3 Price_{iA} \times Income_i$$

$$V_{iB} = \beta_{0B} + \beta_{1B} Income_i + \beta_2 Price_{iB} + \beta_3 Price_{iB} \times Income_i$$

$$\begin{aligned} V_{iA} - V_{iB} = & (0 - \beta_{0B}) \\ & + (0 - \beta_{1B})\, Income_i \\ & + \beta_2 (Price_{iA} - Price_{iB}) \\ & + \beta_3 Income_i \times (Price_{iA} - Price_{iB}) \end{aligned}$$

Putting It All Together: Systematic Utility V_{ij}

- We can set some parameters to zero to obtain identification

$$V_{iA} = 0 + 0 \times Income_i + \beta_2 Price_{iA} + \beta_3 Price_{iA} \times Income_i$$

$$V_{iB} = \beta_{0B} + \beta_{1B} Income_i + \beta_2 Price_{iB} + \beta_3 Price_{iB} \times Income_i$$

- How do we interpret the remaining intercept for Brand B?
  - Relative to the intercept of Brand A
  - If positive, baseline utility of B higher than baseline utility of A
- How do we interpret the income coefficient for Brand B?

Putting It All Together

- How can we model the choice process of customers?
- How to construct a utility function

$$U_{ij} = V_{ij} + \epsilon_{ij}$$

$$V_{ij} = V_j(x_{ij}) + V_j(z_i) + V_j(x_{ij}, z_i) + bias_j$$

- consumer i
- alternative j

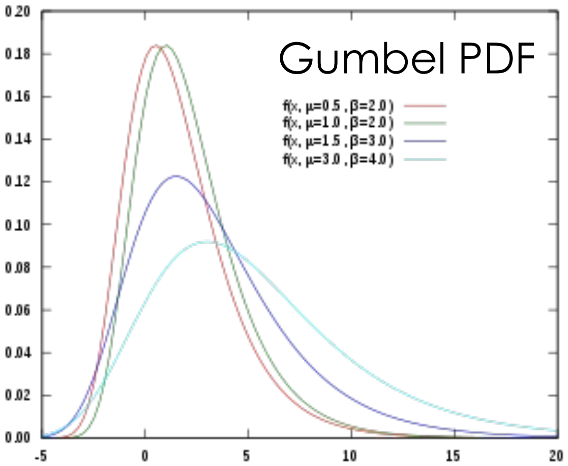
- $V_j(x_{ij})$ is the portion that is associated with the attributes of alternative j faced by consumer i
- $V_j(z_i)$ is the portion of utility associated with the characteristics of the consumer
- $V_j(x_{ij}, z_i)$ contains the interactions between the attributes and characteristics
- $bias_j$ is the alternative specific constant

Choice Models

Stochastic Error Term

Stochastic Part

- We assume that the stochastic part ϵ_{ij} varies across alternatives j and across consumers i
- As the errors are not known, we assume that these come from a probability distribution
- Different assumption on probability distributions of errors leads to different discrete choice model
  - Errors are Gumbel means we get logit models
    - Computational advantages
    - Closed-form choice model
    - Closely approximates normal distribution
    - Gumbel PDF: $f(x) = e^{-(x+e^{-x})}$
  - Errors are normal means we get probit models



Multinomial Logit (MNL) Model

- Since errors ϵ_{ij} are i.i.d. extreme value (Gumbel), we have logit models

  Cumulative density function: $F(x) = e^{-e^{-x}}$ $(\Pr(\epsilon < x) = e^{-e^{-x}})$

- i.i.d means that the errors are independent
  - Across utility equations for a given consumer: $\epsilon_{ij} \perp \epsilon_{ik}$
  - Across different consumers: $\epsilon_{ij} \perp \epsilon_{mj}$ & $\epsilon_{ij} \perp \epsilon_{mk}$
- After estimation, the probability of consumer i choosing brand j is given by

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_k \exp(V_{ik})}$$

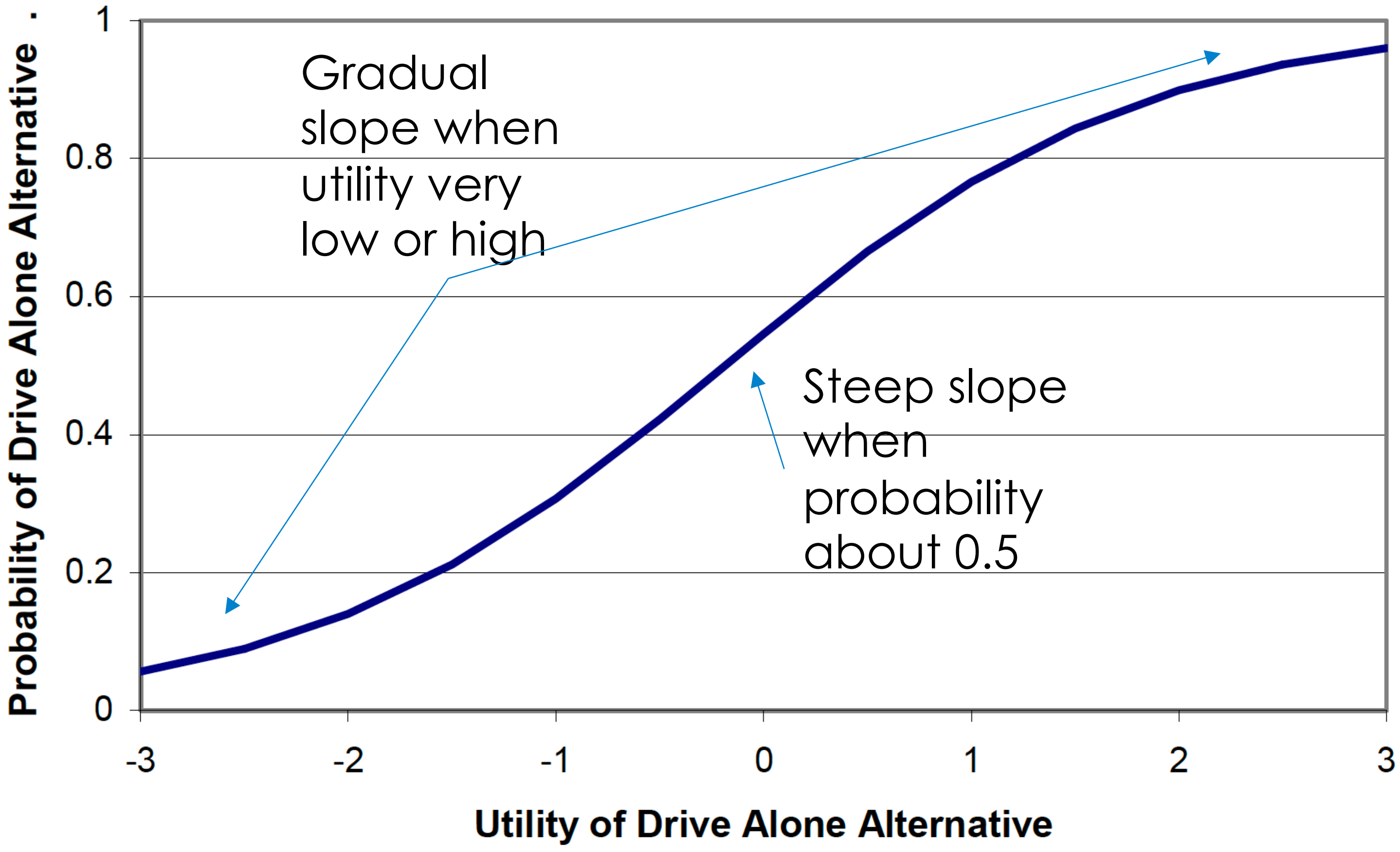
- V_{ij} = systematic part of utility
- k captures all possible alternatives in choice set including j

- Does the above equation for probability make sense?

Multinomial Logit (MNL) Model

Assuming choice is drive alone or not drive alone (i.e., 2 choices)

Probability between 0 and 1



Probability monotonically increasing with utility

Extra – Multinomial Logit Derivation

Don't worry. This will not be tested on the exam.

Probability consumer purchases k vs j:

$$\Pr\left(U_{ik} > U_{ij}\right) = \Pr\left(V_{ik} + \epsilon_{ik} > V_{ij} + \epsilon_{ij}\right)$$

$$= \Pr(V_{ik} - V_{ij} + \epsilon_{ik} > \epsilon_{ij})$$

Imposing CDF of Gumbel distribution and conditioning on ϵ_{ik}, we get:

$$\Pr\left(U_{ik} > U_{ij} | \epsilon_{ik}\right) = \mathrm{F}(V_{ik} - V_{ij} + \epsilon_{ik})$$

Cumulative density function: $\Pr(\epsilon < x) = F(x)$

What about when we have more than alternatives j and k? We can multiply the above probability since the unobserved utility is independent across goods. For all alternatives $j \neq k$,

$$\Pr\left(U_{ik} > U_{ij}\ \forall\, j \neq k | \epsilon_{ik}\right) = \prod_{j \neq k} \mathrm{F}(V_{ik} - V_{ij} + \epsilon_{ik})$$

The unconditional probability (integrating over ϵ_{ik}) that k is chosen is:

$$\mathrm{P}_{\mathrm{ik}} = \Pr\left(U_{ik} > U_{ij}\ \forall\, j \neq k\right) = \int_{-\infty}^{\infty} \prod_{j \neq k} \mathrm{F}\left(V_{ik} - V_{ij} + \epsilon_{ik}\right) f(\epsilon_{ik}) d\epsilon_{ik}$$

From last slide (unconditional probability of choosing k):

$$\mathrm{P_{ik}} = \Pr\left(U_{ik} > U_{ij}\ \forall\, j \neq k\right) = \int_{-\infty}^{\infty} \prod_{j \neq k} \mathrm{F}\left(V_{ik} - V_{ij} + \epsilon_{ik}\right) f(\epsilon_{ik}) d\epsilon_{ik}$$

Imposing Gumbel distribution, we get

Cumulative density function: $F(x) = e^{-e^{-x}}$

$$P_{ik} = \int_{-\infty}^{\infty} \prod_{j \neq k} \exp\left(-\exp\left(-\left(V_{ik} - V_{ij} + \epsilon_{ik}\right)\right)\right) \exp(-\epsilon_{ik}) \exp(-\exp(-\epsilon_{ik}))\, d\epsilon_{ik}$$

Probability density function: $f(x) = e^{-(x+e^{-x})}$

Since $\exp(-\exp(-\epsilon_{ik})) = \exp(-\exp(-(V_{ik} - V_{ik} + \epsilon_{ik})))$, the above simplifies to:

$$P_{ik} = \int_{-\infty}^{\infty} \prod_{j} \exp\left(-\exp\left(-\left(V_{ik} - V_{ij} + \epsilon_{ik}\right)\right)\right) \exp(-\epsilon_{ik})\, d\epsilon_{ik}$$

(Product now includes k)

Since product of exponentials is the exponential of the sum of the exponents:

$$P_{ik} = \int_{-\infty}^{\infty} \exp\left(-\sum_{j} \exp\left(-\left(V_{ik} - V_{ij} + \epsilon_{ik}\right)\right)\right) \exp(-\epsilon_{ik})\, d\epsilon_{ik} = \int_{-\infty}^{\infty} \exp\left(-\exp(-\epsilon_{ik}) \sum_{j} \exp\left(-\left(V_{ik} - V_{ij}\right)\right)\right) \exp(-\epsilon_{ik})\, d\epsilon_{ik}$$

We need to use a change of variables:

$$t = -\exp(-\epsilon_{ik})$$
$$dt = \exp(-\epsilon_{ik})\, d\epsilon_{ik} \text{ where } t \in (-\infty, 0)$$

Then,

$$P_{ik} = \int_{-\infty}^{\infty} \exp\left(-\exp(-\epsilon_{ik}) \sum_j \exp\left(-(V_{ik} - V_{ij})\right)\right) \exp(-\epsilon_{ik})\, d\epsilon_{ik} = \int_{-\infty}^{0} \exp\left(t \sum_j \exp\left(-(V_{ik} - V_{ij})\right)\right) dt$$

Completing the integral:

$$P_{ik} = \left(\frac{\exp\left(t \sum_j \exp\left(-(V_{ik} - V_{ij})\right)\right)}{\sum_j \exp\left(-(V_{ik} - V_{ij})\right)}\right) \Big|_{-\infty}^{0}$$

$$P_{ik} = \frac{1}{\sum_j \exp\left(-(V_{ik} - V_{ij})\right)} = \frac{1}{\sum_j \exp\left(-(V_{ik})\right) \exp(V_{ij})} = \frac{1}{\exp(-V_{ik}) \sum_j \exp\left(V_{ij}\right)} = \frac{\exp(V_{ik})}{\sum_j \exp\left(V_{ij}\right)}$$

Break

10-minutes

Choice Models

Estimation

So What? From Theory to Practice!

- We established a framework for how people make choices **BUT** how do we use it?
- In practice, we want to estimate the coefficients in systematic part of the utility
  - Why? For interpretation!
  - How price sensitive are consumers?
  - How time sensitive are consumers?
  - When buying a car, how sensitive are consumers to mileage per gallon?

So What? From Theory to Practice!

- Consider data with 200 consumer choices:
  - Two brands, named 1 and 2
  - One alternative specific variable named Price
- What are the systematic utility functions for the two brands?
  - Systematic utility for consumer i choosing alternative 1: $V_{i1} = \beta Price_{i1} + \beta_{01}$
  - Systematic utility for consumer i choosing alternative 2: $V_{i2} = \beta Price_{i2} + \beta_{02}$
- Based on the theory, how many parameters can we estimate?
  - 2
  - Systematic utility for consumer i choosing alternative 1: $V_{i1} = \beta Price_{i1}$
    - Intercept (β_{01})= 0
  - Systematic utility for consumer i choosing alternative 2: $V_{i2} = \beta Price_{i2} + \beta_{02}$

Data Extract

- Here is an extract from 9 consumers

Consumer	Choice	Price_1	Price_2
1	1	1.15	1.2
2	1	1.15	1.24
3	2	1.1	1.09
4	2	1.15	1.2
5	2	1.1	0.9
6	2	1.1	1.24
7	2	1.1	1.2
8	2	1.15	0.9
9	1	1.25	1.2

- We can use data to estimate the intercept for brand 2 and the price coefficient
- How?

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) – Reminder

- Maximum Likelihood Estimation
  - Use the data to find values of the model parameters (θ) that maximize the likelihood of observing the data that we have

- We estimate the model parameters by maximizing the likelihood function $L(\theta)$

- The resulting parameter estimates, θ_{ML} are called "maximum likelihood estimates"

Likelihood – Consumer Choice

- Let y_i be the observed choice for customer i, which takes the values 1 or 2
- Let δ_{i1} and δ_{i2} be two binary variables
  - $\delta_{ij} = 1$, if $y_{ij} = j$, and zero otherwise
- The likelihood for observation i is given by

$$\mathcal{L}_i(\theta|\, y_i\,) = \text{Prob}(y_i = 1)^{\delta_{i1}}\text{Prob}(y_i = 2)^{\delta_{i2}}$$

  - Why? Note that $\delta_{i2} = 1 - \delta_{i1}$ so δ_{ij} serves as an indicator variable for probability
- The likelihood for the entire data is the product of the observation-level likelihoods

$$\mathcal{L}(\theta|D) = \prod_{i=1}^{N} \mathcal{L}_i(\theta|y_i)$$

Log-Likelihood

Recall probability of consumer i choosing brand j,

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_k \exp(V_{ik})}$$

- In practice, we maximize the log-likelihood
- The log-likelihood for consumer i is given by

$$\mathcal{LL}_i(\theta|y_i) = \delta_{i1}\log\big(\text{Prob}(y_i = 1|\theta)\big) + \delta_{i2}\log\big(\text{Prob}(y_i = 2|\theta)\big) =$$

$$\delta_{i1}\log\left(\frac{\exp(V_{i1})}{\sum_k \exp(V_{ik})}\right) + \delta_{i2}\log\left(\frac{\exp(V_{i2})}{\sum_k \exp(V_{ik})}\right)$$

- The overall log-likelihood is the sum of the consumer-specific log-likelihoods

$$\mathcal{LL}(\theta|D) = \sum_{i=1}^{N} \mathcal{LL}_i(\theta|y_i)$$

Your Turn!

- Open dataBinaryAnalysisProblems.xlsx
- Start with an intercept and a β_{price} of 0.5
- For each individual (each row), add columns that compute
  - The systematic utilities for each brand
  - The probabilities of purchasing each brand
  - The loglikelihood (you can use more columns if you prefer)
    - $\mathcal{LL}_i(\theta|y_i) = \delta_{i1} \log\left(\frac{\exp(V_{i1})}{\sum_k \exp(V_{ik})}\right) + \delta_{i2} \log\left(\frac{\exp(V_{i2})}{\sum_k \exp(V_{ik})}\right)$
- Add a cell that contains the total log-likelihood and use solver to find the coefficients
  - $\mathcal{LL}(\theta|D) = \sum_{i=1}^{N} \mathcal{LL}_i(\theta|y_i)$
- You have 20 minutes!

In-Class Exercise Solution

- Systematic utility for consumer i choosing alternative 1: $V_{i1} = \beta Price_1$
  - Intercept (β_{01})= 0
- Systematic utility for consumer i choosing alternative 2: $V_{i2} = \beta Price_2 + \beta_{02}$

- $\mathcal{LL}(\theta|D) = \sum_{i=1}^{N} \delta_{i1} \log\left(\frac{\exp(V_{i1})}{\exp(V_{i1})+\exp(V_{i2})}\right) + \delta_{i2} \log\left(\frac{\exp(V_{i2})}{\exp(V_{i1})+\exp(V_{i2})}\right)$

β_{02}	β
0.434886855	**-1.2880701**
Total LogLikelihood	-131.201885

Choice Models

Model Comparison

MNL Application - Megabus

- Data: 210 travelers stated their choice among 4 travel modes (1=Air, 2=Bus, 3=Car, 4=Train)
- Independent variables
  - Time: terminal waiting time
  - Invc: In-vehicle cost
  - Invt: In-vehicle time
  - Hinc: Household income in thousands
- We normalize by dividing each of these variables by 100 in the data, before estimation
  - Note: We could have standardized instead
- You will work on this dataset for the in-class concept check

MNL Travel Data Estimates – Model 1

Parameter	Estimate	Std Error	2.5%	97.5%
intBus	-1.434	0.681	-2.768	-0.099
intCar	-4.740	0.868	-6.440	-3.040
intTrain	-0.787	0.603	-1.968	0.394
Time	-9.689	1.034	-11.716	-7.662
Invc	-1.391	0.665	-2.695	-0.088
Invt	-0.400	0.085	-0.566	-0.233

95% Confidence Interval

- Intercept for plane = 0
- Only attributes of options included
- No consumer characteristics

- Time: terminal waiting time,
- Invc: In-vehicle cost
- Invt: In-vehicle time

MNL Travel Data Estimates – Model 2

Parameter	Estimate	Std Error	2.5%	97.5%
intBus	-0.184	0.897	-1.942	1.573
intCar	-4.247	1.007	-6.220	-2.275
intTrain	1.242	0.817	-0.359	2.843
Time	-9.528	1.036	-11.558	-7.499
Invc	-0.450	0.721	-1.863	0.964
Invt	-0.366	0.087	-0.537	-0.196
Bus_Hinc	-2.311	1.646	-5.537	0.914
Car_Hinc	0.210	1.210	-2.160	2.581
Train_Hinc	-5.590	1.536	-8.600	-2.580

- Intercept for plane = 0
- Plane_Hinc = 0

- Time: terminal waiting time,
- Invc: In-vehicle cost

- Invt: In-vehicle time
- Hinc: Household income in thousands

Model Comparison

- Models can be compared using the Bayes Information Criterion (BIC)

- BIC is given by

$$\text{BIC} = -2 * LL(\theta_{ML}) + K * \ln(N)$$

  - K is the total number of parameters estimated
    - BIC penalizes having more parameters
  - N is the total number of observations

- Lower BIC = better. Why?

- Model 1: $\text{BIC\_1} = -2 * (-192.89) + 6 * \ln(210) = 417.86$
- Model 2: $\text{BIC\_2} = -2 * (-182.22) + 9 * \ln(210) = 412.564$

Elasticities & IIA

Choice Probability Derivatives

- Probabilities are functions of observed variables
- By varying variables, we can analyze how probabilities vary
  - How does an increase in price impact purchase probability?
  - How? derivatives
- Own Derivatives
  - To what extent will the probability of choosing Bus change when we decrease Bus's cost?
- Cross Derivatives
  - To what extent will the probability of choosing Bus change when we decrease Train's price?

Own Derivatives

- Let i be consumer, j be alternative, and m be an attribute (e.g., price)
- P_{ij} is the probability that consumer i chooses j
- Own Derivative: Impact on probability of alternative j when attribute of j is changed

$$\frac{\partial P_{ij}}{\partial x_{ijm}} = \beta_m P_{ij}(1 - P_{ij})$$

- Notice that P_{ij} is function of x_{ijm}. How?
- When is the derivative null?
  - When $P_{ij} = 0$ or $P_{ij} = 1$: no uncertainty in purchase choice
- When is the derivative the highest?
  - When $P_{ij} = 0.5$: uncertainty is the highest

Own Elasticities

- Own choice elasticity is given by

$$\frac{\%\ \text{Change in Probability of } j}{\%\ \text{Change in an attribute of } j}$$

- Own elasticity is

$$E_{jx_{ijm}} = \frac{\frac{\partial P_{ij}}{P_{ij}}}{\frac{\partial x_{ijm}}{x_{ijm}}} = \frac{\partial P_{ij}}{\partial x_{ijm}} \frac{x_{ijm}}{P_{ij}} = \beta_m P_{ij}(1 - P_{ij}) \left(\frac{x_{ijm}}{P_{ij}}\right) = \beta_m x_{ijm}(1 - P_{ij})$$

How important the attribute is (β_m)

Value of the attribute (x_{ijm})

Probability of choosing the alternative (P_{ij})

Cross Derivatives

- Let i be consumer, j and k be alternatives, and m be an attribute (e.g., price)
- Cross Derivative: Impact on probability of alternative *j* when attribute of *k* is changed

$$\frac{\partial P_{ij}}{\partial x_{ikm}} = -\beta_m P_{ij} P_{ik}$$

- When is the above derivative the highest?
    - Unclear

Cross Elasticities

- Cross Choice elasticities is given by

$$\frac{\%\text{ Change in Probability of } j}{\%\text{ Change in an attribute of } k}$$

- Cross elasticity for j is

$$E_{jx_{ikm}} = \frac{\partial P_{ij}}{\partial x_{ikm}} \frac{x_{ikm}}{P_{ij}} = -\beta_m x_{ikm} P_{ik}$$

- Cross elasticities are the same for all j
  - When k changes its attribute value by 1 percent, it impacts the probabilities of all other alternatives by the same percentage

Elasticities

- Own and Cross elasticities with respect to Invt (In Vehicle Time)
- Effect on the choice probability of the row alternative when the time of the column alternative changes

	Air	Bus	Car	Train
Air	-0.252	0.019	0.023	0.039
Bus	0.013	-0.094	0.023	0.039
Car	0.013	0.019	-0.022	0.039
Train	0.013	0.019	0.023	-0.100

- Notice cross elasticities in each column
  - This pattern is due to Independence of Irrelevant alternatives

Independence of Irrelevant Alternatives

- Independence of Irrelevant Alternatives (IIA)
  - Ratio of choice probabilities between pairs of alternatives is independent of availability or attributes of other alternatives

$$\frac{P_{ij}}{P_{ik}} = \frac{\exp(V_{ij})}{\exp(V_{ik})}$$

- Characteristics of one particular choice alternative do not impact the relative probabilities of choosing other alternatives
- Why?
- Denominator is the same for all probabilities and numerator only depends on alternative

Independence of Irrelevant Attributes

- Suppose consumers are indifferent between a Car and a Red Bus.
- Then $P(car) = 0.5$, $P(RedBus) = 0.5$

$$\frac{P(car)}{P(RedBus)} = 1$$

- The company introduces a Blue Bus: identical to the Red Bus, except for its color (irrelevant attribute)
- → Utilities for car and Red bus don't change
- According to MNL $\frac{P(car)}{P(RedBus)} = 1$ and we expect that $\frac{P(RedBus)}{P(BlueBus)} = 1$
- What are probabilities of all the alternatives?

What are probabilities of all the alternatives?

0

$P\,(car)$=0.33, $P\,(RedBus)$=0.33, and $P\,(BlueBus)$=0.33

0

$P\,(car)$=0.5, $P\,(RedBus)$=0.25, and $P\,(BlueBus)$=0.25

0

Independence of Irrelevant Attributes

- Suppose consumers are indifferent between a Car and a Red Bus.
- Then $P(car) = 0.5$, $P(RedBus) = 0.5$

$$\frac{P(car)}{P(RedBus)} = 1$$

- The company introduces a Blue Bus: identical to the Red Bus, except for its color (irrelevant attribute) → Utilities for car and Red bus don't change

- According to MNL $\frac{P(car)}{P(RedBus)} = 1$ and we expect that $\frac{P(RedBus)}{P(BlueBus)} = 1$

- What are probabilities of all the alternatives?
  - $P(car) = 0.33$, $P(RedBus) = 0.33$, and $P(BlueBus) = 0.33$

Blue Bus - Red Bus

- MNL implies the following choice probabilities

	Car	Red Bus	Blue Bus
Two Alternatives	0.5	0.5	NA
MNL: Three Alternatives	0.33	0.33	0.33
Expected: Three Alternatives	0.5	0.25	0.25

- The new alternative draws proportionally from each of the existing alternatives

IIA

- IIA is beneficial in modeling when choice sets differ across observations
  - Allows addition or removal of an alternative from choice set
  - Why? Structure and parameters of the model won't be impacted

- But
  - Can be problematic in predicting choice shares when new brands are introduced
  - Can give misleading elasticities
  - Alternatives?
    - Nested Logit; Multinomial Probit…

Concept Check In-Class

Choice Modeling

Model 1 – Systematic Utilities

- Elements of model: intercept, time, in-vehicle cost, in-vehicle time
- How many parameters can be identified?
  - 6 (3 intercepts, time, invc, invt)

- Systematic utility for consumer i choosing air:

$$V_{i,air} = \beta_{time} Time_{i,air} + \beta_{invc} Invc_{i,air} + \beta_{invt} Invt_{i,air}$$

  - Intercept ($\beta_{0,air}$)= 0
- Systematic utility for consumer i choosing bus:

$$V_{i,bus} = \beta_{bus} + \beta_{time} Time_{i,bus} + \beta_{invc} Invc_{i,bus} + \beta_{invt} Invt_{i,bus}$$

- Similar to above for choosing car or train

Model 1 – Log Likelihood

$$\mathcal{LL}(\theta|D) = \sum_{i=1}^{N} \begin{matrix} \delta_{i,air} \log\left(\dfrac{\exp(V_{i,air})}{\exp(V_{i,air}) + \exp(V_{i,bus}) + \exp(V_{i,car}) + \exp(V_{i,train})}\right) + \\ \delta_{i,bus} \log\left(\dfrac{\exp(V_{i,bus})}{\exp(V_{i,air}) + \exp(V_{i,bus}) + \exp(V_{i,car}) + \exp(V_{i,train})}\right) + \cdots \end{matrix}$$

β_{bus}	β_{car}	β_{train}	β_{time}	β_{invc}	β_{invt}
-1.43371	-4.73997	-0.78674	-9.68878	-1.3912	-0.39947

Total LogLike

-192.889

Model 1 – BIC

- BIC is given by

$$\text{BIC} = -2 * LL(\theta_{ML}) + K * \ln(N)$$

  - K is the total number of parameters estimated – **6**
  - N is the total number of observations – **210**

- **BIC = 417.86**

Model 2 – Systematic Utilities

- Elements of model: intercept, time, in-vehicle cost, in-vehicle time, **household income**
- How many parameters can be estimated?
  - 9 (3 intercepts, 3 household income, time, invc, invt)

- Systematic utility for consumer i choosing air:

$$V_{i,air} = \beta_{time} Time_{i,air} + \beta_{invc} Invc_{i,air} + \beta_{invt} Invt_{i,air}$$

  - Intercept ($\beta_{0,air}$)= 0, $\beta_{air,hinc} = 0$
- Systematic utility for consumer i choosing bus:

$$V_{i,bus} = \beta_{bus} + \beta_{time} Time_{i,bus} + \beta_{invc} Invc_{i,bus} + \beta_{invt} Invt_{i,bus} + \beta_{bus,hinc} Hinc_i$$

  - Similar to above for choosing car or train

Model 2 – Prediction for Individual 1

β_{bus}	β_{car}	β_{train}	β_{time}	β_{invc}	β_{invt}	$\beta_{bus,hinc}$	$\beta_{car,hinc}$	$\beta_{train,hinc}$
-0.1844	-4.2476	1.2421	-9.5285	-0.4499	-0.3665	-2.3111	0.2103	-5.5896

Id	Choice	Time.air	Invc.air	Invt.air	Time.bus	Invc.bus	Invt.bus
1	car	0.69	0.59	1	0.35	0.25	4.17

Time.car	Invc.car	Invt.car	Time.train	Invc.train	Invt.train	Hinc
0	0.1	1.8	0.34	0.31	3.72	0.35

Id	Vair	Vbus	Vcar	Vtrain	Den	Prob(air)	Prob(bus)	Prob(car)	Prob(Train)
1	-7.20663	-5.96901	-4.87866	-5.45675	0.015173	0.04888	0.168508	0.501364	0.281248

Highest probability = car

Model 2 – Elasticities

- Own choice elasticity is given by

$$\frac{\%\text{ Change in Probability of } j}{\%\text{ Change in an attribute of } j}$$

- Own elasticity is

How important the attribute is

Value of the attribute

$$E_{air,invt} = \beta_{invt} invt_{1,air}(1 - P_{1,air}) = -0.36647 * 1(1 - 0.04888)$$

$$= -0.349$$

Probability of choosing the alternative

Interpretation
A 10% increase in in-vehicle time for air travel reduces the choice probability of flying by 3.49% for Individual 1

Let's Go to Python

Choice Modeling

Takeaways

- Multinomial Logit is the most widely used choice model
- Identification: Only differences in utility matter
  - Need to set one alternative specific constant to zero
  - Need to set the coefficients of the individual characteristics to zero for one alternative
- Can be used to predict brands bought on different purchase occasions
- Can be used to compute own and cross-elasticities
- Beware of IIA

Break

5-minutes

Midterm Review

Week 1 – Marketing Datasets

- Each dataset has pros and cons
  - What are they?
  - What type of question can I answer?

- Important to quickly know what is possible or not with your data

Data Taxonomy

	Primary Data *Data that is gathered by the researcher for the purpose of answering a specific question.*	**Secondary Data** *Data that was gathered for a purpose other than answering the specific question.*
Structured *Data that can be easily and meaningfully represented and manipulated in a traditional database (spreadsheet). Typically numeric or "choice" data.*	Surveys (ratings, choice) Experiments	Transaction logs Scanner panel data Ad tracking Product usage data
Unstructured *Data that cannot be meaningfully stored in a traditional data structure (spreadsheet) without further processing. Examples include text, images, video, and voice.*	Focus groups Interviews Surveys (free response) Observation Eye tracking Physiological/neural	Online reviews Social media Most digital content Call logs

Types of Marketing Research

Exploratory Research

(Ambiguous Problem)

"Our sales are declining and we do not know why."

Descriptive Research

(Aware of Problem)

"What kinds of people are buying our products?"

"Who buys our competitors' products?"

Causal Research

(Problem Clearly Defined)

"Will buyers purchase more of our product in a new package?"

Week 2 – Segmentation and Targeting

- What is STP?

*Deliver the **right products**, to the **right people**, in the **right way***

Targeting Segmentation Positioning

- What type of data can we use for segmentation?
  - Geodemographics, psychographics, behavioral, benefits and needs
- How to implement and interpret results from hierarchical clustering and k-means
  - Basic idea: use similarity in columns to group rows in segments
  - Hierarchical clustering: sequentially join individuals together based on distance until we get one large unique cluster then select number of segments
  - K-means: find groups of data that are the same within and distinct across groups
- You should be able to determine the number of segments and interpret them
- Segments are **L**arge, **I**dentifiable, **D**istinctive, **S**table and **actionable!**
- How to choose a target segment? Opportunity + Competition + Customer + Company "fit"

Week 3 – Segmentation and Positioning

- Dimension reduction techniques – Factor Analysis (PCA)
  - Assume that independent variables are derived from underlying "concepts"
  - Uncover underlying structure between many variables
- Steps to PCA: determining the number of factors and interpreting them
  - Good factors: uncorrelated, capture as much of the original variance as possible
  - Factors are often intuitive, easier to use, and managerially interesting
- Understand the difference between loading, communalities and scores
  - Loadings = how the original variables relate to the factors
  - Communalities = how much variability in the original variables is explained by the factors
  - Scores = translation of original data into factors

Week 4 – Perceptual Maps + Ford Ka + Customer Lifetime Value (CLV)

- Be able to build, interpret and use a perceptual map
- How to conduct an end-to-end marketing strategy
  - Be aware of common potential problems: how to reach target, data limitations,…
- Margin *m*; Retention rate *r*; Discount rate *i*; Acquisition cost *AC*

$$CLV = m\left(\frac{r}{1+i-r}\right) - AC$$

- Understand the impact of each parameter on CLV
  - E.g., How much should a company spend to acquire a new account?

Week 5 – CRM + Churn

- Managing the CLV
  - Customer acquisition
    - Acquisitions, affiliation network…
  - Customer expansion
    - Bundling, recommendation (matrix factorization),…
  - Customer retention
    - Causes of churn, double effect of high retention, impact on market share, firm value…
- Collaborative filtering for recommendation systems
- Discrete survival models for estimating customer churn
  - Be able to replicate the logic
    - Geometric
    - Finite mixture model

Questions?