Choice data

Michel Bierlaire

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

Outline

1. Introduction
2. Revealed preferences
3. Stated preferences
4. Rp vs SP
5. Experimental design

Sampling

- Identify the population of interest.
- In general, it is not possible to collect data about each individual.
- Identify a list of N representative individuals.
- Various sampling methods are presented later in this course.
- Collect choice data for each individual in the sample.

Choice context

Revealed preferences

- Observe actual behavior.
- Real market situations.
- Example: scanner data in supermarkets.

Stated preferences

- Hypothetical situations.
- Choice context defined by the analyst.
- Example: Swissmetro.

Revealed preferences

Data about the decision-maker: socio-economic characteristics

- Age, income, level of education, etc.
- Collected in any survey.
- Not specific to choice models.
- Collect those that seem relevant for the analysis.

Choice set

- Identify the list of alternatives considered by the respondent.
- Context dependent.
- Awareness difficult to observe.

Revealed preferences

Data about the alternatives

- Utility is a latent concept, cannot be observed.
- Value of the attributes.
- Particularly difficult for non chosen alternatives.

Observed outcome

- The chosen alternative

Stated preferences

Hypothetical situations

- Choice context is constructed by the analyst.
- Several scenarios can be created for each respondent.
- Preferences are expressed through statements or intentions.

Stated preferences

Data about the decision-maker: socio-economic characteristics

- Age, income, level of education, etc.
- Collected in any survey.
- Not specific to choice models.
- Collect those that seem relevant for the analysis.

Choice set

- Constructed by the analyst.
- May contain hypothetical alternatives.
- May vary across scenarios and across respondents.

Stated preferences

Data about the alternatives

- Constructed by the analyst.
- Provided for each alternative
- Experimental design.

Preferences

- Choice
- Ranking
- Rating
- Allocation

Preference data

Consider the following beers

1. Cardinal
2. Kronenbourg
3. Orval
4. Tsing Tao

Choice

What is your preferred option?

Preference data

Consider the following beers

1. Cardinal
2. Kronenbourg
3. Orval
4. Tsing Tao

Ranking

Rank the beers, from the best to the worst

Preference data

Consider the following beers

1. Cardinal
2. Kronenbourg
3. Orval
4. Tsing Tao

Rating

Associate a rate from 0 (worst) to 10 (best) with each beer

Preference data

Consider the following beers

1. Cardinal
2. Kronenbourg
3. Orval
4. Tsing Tao

Allocation

Distribute 100 points among the beers

Ranking

Pros

- More info than the choice

Cons

- Best and worse easy, others more arbitrary
- Analyst cannot distinguish between real preference and random order
- Possible inconsistencies

Rating

Pros

- Concept close to utility
- More information than ranking

Cons

- Difficult task
- Scale is arbitrary
- Scale is person specific: two individuals with the same preferences may give a different scale
- Scale depends on history: if B is rated after A, its rate will depend on the rate of A

Allocation

Pros

- Concept close to market shares
- Scale is normalized

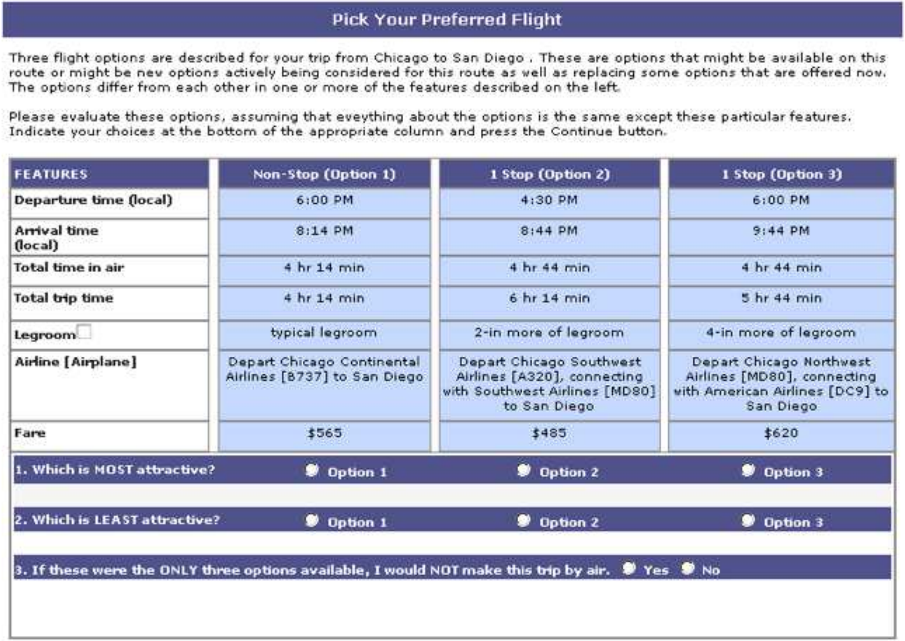
Cons

- Abstract task
- Two individuals with the same preferences may give a different scale
- Artificial emphasis on 0% and 100%
- Rounding issues

Example

Boeing Commercial Airplanes

- 2004—2005.
- Designed by Boeing staff with the assistance of Jordan Louviere of the University of Technology, Sydney.
- Objective: understanding the sensitivity that air passengers have toward the attributes of an airline itinerary.
- Recruitment: intercepting customers of an internet airline booking service that searches for low-cost travel deals



Pick Your Preferred Flight

Three flight options are described for your trip from Chicago to San Diego . These are options that might be available on this route or might be new options actively being considered for this route as well as replacing some options that are offered now. The options differ from each other in one or more of the features described on the left.

Please evaluate these options, assuming that eveything about the options is the same except these particular features. Indicate your choices at the bottom of the appropriate column and press the Continue button.

FEATURES	Non-Stop (Option 1)	1 Stop (Option 2)	1 Stop (Option 3)
Departure time (local)	6:00 PM	4:30 PM	6:00 PM
Arrival time (local)	8:14 PM	8:44 PM	9:44 PM
Total time in air	4 hr 14 min	4 hr 44 min	4 hr 44 min
Total trip time	4 hr 14 min	6 hr 14 min	5 hr 44 min
Legroom □	typical legroom	2-in more of legroom	4-in more of legroom
Airline [Airplane]	Depart Chicago Continental Airlines [B737] to San Diego	Depart Chicago Southwest Airlines [A320], connecting with Southwest Airlines [MD80] to San Diego	Depart Chicago Northwest Airlines [MD80], connecting with American Airlines [DC9] to San Diego
Fare	$565	$485	$620

1. Which is MOST attractive? ○ Option 1 ○ Option 2 ○ Option 3

2. Which is LEAST attractive? ○ Option 1 ○ Option 2 ○ Option 3

3. If these were the ONLY three options available, I would NOT make this trip by air. ○ Yes ○ No

RP data: drawbacks

- Limited to existing alternatives, attributes and attributes levels.
- Lack of variability of some attributes
- Lack of information about non chosen alternatives
- High level of correlation
- Data collection cost
- In general, one individual = one observation

SP data: advantages

- Exploring new alternatives, attributes and attributes levels
- Control of the attributes variability
- Control on all alternatives
- Control on the level of correlation
- One individual can answer several questions

SP data: drawbacks

- Hypothetical situations
- Cannot be used for market shares
- Decision-makers do not have to assume their choice
- “A bike or a Ferrari?” — “A Ferrari, of course!”
- Real constraints not involved
- Credibility
- Valid within the range of the experimental design
- Policy bias (example: “every body else should take the bus”)
- Justification bias (or inertia)
- Framing: phrasing of the question matters
- Anchoring: one variable explains it all
- Fatigue effect

Experimental design

Experiment

An experiment is a set of actions and observations, performed to verify or falsify a hypothesis or research a causal relationship between phenomena. The design of the experiment, or *experimental design* is the definition of the set of actions.

Multi-variable experiment

- Dependent variables (e.g. choice) are related to independent variables (travel time,cost, etc.)
- Independent variables are considered at given levels

Experimental design

Example

- Context: Swissmetro between Lausanne and Zürich
- Objective: identify mode share changes with Swissmetro

Definition of the choice set

car as driver, *car as passenger*, train, Swissmetro, *helicopter*, *taxi*

Experimental design

Definition of the list of attributes

- mode-specific:
  - train: frequency, waiting time, fares, etc.
  - car: fuel, toll, parking costs, etc.
- shared by modes:
  - departure time
  - arrival time
  - comfort

Stimuli definition

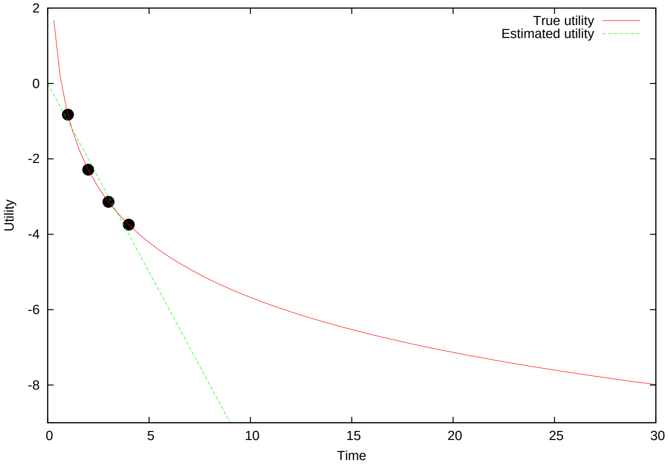
Definition of the levels

numbers or words

Issues

- number of levels?
- range, extreme values
- realism vs. completeness
- Realism: only some values make sense
- Completeness: need sufficient information to estimate the model

Stimuli definition



Stimuli definition

Necessity to explain the meaning of the levels

Example: comfort

- Low: “Hard seats. No air conditioning. No table. No power supply. No internet.”
- Medium: “Soft seats. Air conditioning. Small tables. No power supply. No internet.”
- High: “Soft seats. Air conditioning. Large individual tables. Power supply. Wireless internet.”

Full factorial design

	Comfort	Travel time	Comfort	Travel time
1	Low	30 min	1	1
2	Low	60 min	1	2
3	Low	90 min	1	3
4	Low	120 min	1	4
5	Medium	30 min	2	1
6	Medium	60 min	2	2
7	Medium	90 min	2	3
8	Medium	120 min	2	4
9	High	30 min	3	1
10	High	60 min	3	2
11	High	90 min	3	3
12	High	120 min	3	4

Generation of the design

Orthogonal coding

- Sum up to 0 columnwise
- Only odd numbers are used
- $2k+1$ levels (odd): $\{-2k+1, \ldots -3, -1, 0, 1, 3, \ldots, 2k-1\}$
- $2k$ levels (even): $\{-2k+1, \ldots -3, -1, 1, 3, \ldots, 2k-1\}$

Generation of the design

	Comfort	Travel time	Comfort	Travel time
1	Low	30 min	-1	-3
2	Low	60 min	-1	-1
3	Low	90 min	-1	1
4	Low	120 min	-1	3
5	Medium	30 min	0	-3
6	Medium	60 min	0	-1
7	Medium	90 min	0	1
8	Medium	120 min	0	3
9	High	30 min	1	-3
10	High	60 min	1	-1
11	High	90 min	1	1
12	High	120 min	1	3

Generation of the design

	Train	Swissmetro
Comfort	High	Low
Travel time	120 min	30 min
Choice :	❒	✔

	Train	Swissmetro
Comfort	Low	Medium
Travel time	90 min	60 min
Choice :	✔	❒

	Train	Swissmetro
Comfort	Medium	High
Travel time	60 min	90 min
Choice :	✔	❒

Generation of the design

Curse of dimensionality

- 2 alternatives, 3 levels for comfort, 4 levels for travel time = 24 combinations
- Number of questions grows exponentially
- Necessary to reduce the number

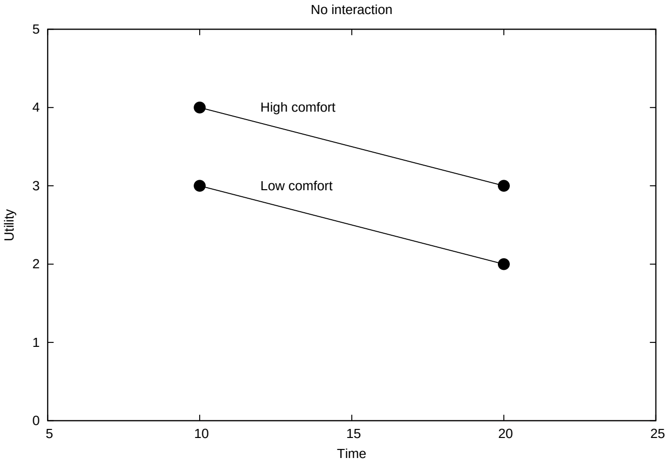
Effects

Main effect

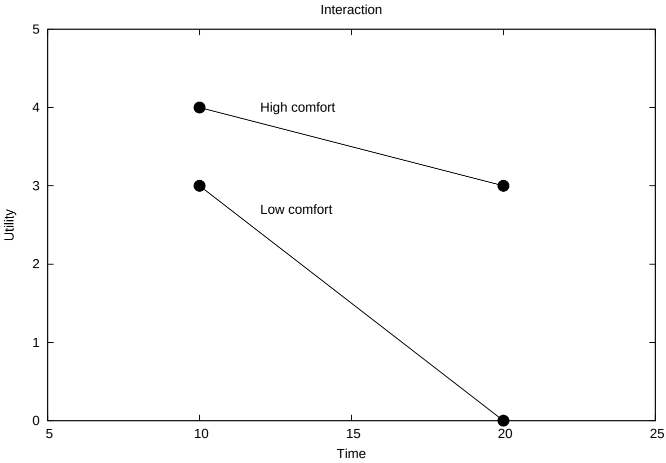
The main effect of a variable is the effect of the experimental response of going from one level of the variable to the next given that the remaining variables do not change

If the effect of two independent variables is not additive, the variables are said to *interact*.

Effects


No interaction
High comfort
Low comfort
Utility
Time
0
1
2
3
4
5
5
10
15
20
25

Effects


Interaction
Utility
Time
5
4
3
2
1
0
5
10
15
20
25
High comfort
Low comfort

Effects

No interaction

$$U = \beta_1 \text{time} + \beta_2 \text{HighComfort}$$

Interaction

$$U = \beta_1 \text{time} + \beta_2 \text{HighComfort} + \beta_3 \text{Time} \cdot \text{HighComfort}$$

Reducing the design

Full factorial design:

	Mode	Comfort	Travel Time
1	Train	Medium	90
2	Train	Medium	120
3	Train	High	90
4	Train	High	120
5	Swissmetro	Medium	90
6	Swissmetro	Medium	120
7	Swissmetro	High	90
8	Swissmetro	High	120

Reducing the design

Coded full factorial design:

	Mode	Comfort	Travel Time
1	-1	-1	-1
2	-1	-1	1
3	-1	1	-1
4	-1	1	1
5	1	-1	-1
6	1	-1	1
7	1	1	-1
8	1	1	1

Reducing the design

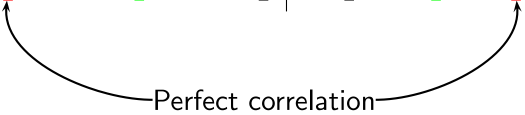
Main effects and interactions

	Mode	Comfort	T. Time	M-C	M-T	C-T	M-C-T
1	-1	-1	-1	1	1	1	-1
2	-1	-1	1	1	-1	-1	1
3	-1	1	-1	-1	1	-1	1
4	-1	1	1	-1	-1	1	-1
5	1	-1	-1	-1	-1	1	1
6	1	-1	1	-1	1	-1	-1
7	1	1	-1	1	-1	-1	-1
8	1	1	1	1	1	1	1

Reducing the design

Fractional factorial design

	Mode	Comfort	T Time	M-C	M-T	C-T	M-C-T
2	-1	-1	1	1	-1	-1	1
3	-1	1	-1	-1	1	-1	1
5	1	-1	-1	-1	-1	1	1
8	1	1	1	1	1	1	1


Perfect correlation

Impossible to distinguish between C-T and mode.

Reducing the design

In practice...

- It is critical to capture main effects
- Three-way interactions (and higher) can be ignored
- Important to choose only a few two-way interactions to be captured
- Compute the correlation matrix of the design to identify confounding effects

Generation of the design

Blocking

- Divide the design into blocks
- Give a different block to different individuals
- Use a blocking attribute orthogonal to the design
- Example: use the 3-way interaction variable in the example above

Reducing the design

Blocks: 3-way interactions are biased

	Mode	Comf.	T Time	M-C	M-T	C-T	M-C-T	Block
1	-1	-1	-1	1	1	1	-1	-1
2	-1	-1	1	1	-1	-1	1	1
3	-1	1	-1	-1	1	-1	1	1
4	-1	1	1	-1	-1	1	-1	-1
5	1	-1	-1	-1	-1	1	1	1
6	1	-1	1	-1	1	-1	-1	-1
7	1	1	-1	1	-1	-1	-1	-1
8	1	1	1	1	1	1	1	1
	0	0	0	0	0	0	8	

Reducing the design

Blocks: mode and 3-way interactions are biased

	Mode	Comf.	T Time	M-C	M-T	C-T	M-C-T	Block
1	-1	-1	-1	1	1	1	-1	-2
2	-1	-1	1	1	-1	-1	1	1
3	-1	1	-1	-1	1	-1	1	1
4	-1	1	1	-1	-1	1	-1	-2
5	1	-1	-1	-1	-1	1	1	2
6	1	-1	1	-1	1	-1	-1	-1
7	1	1	-1	1	-1	-1	-1	-1
8	1	1	1	1	1	1	1	2
	4	0	0	0	0	0	12	

Conclusion

- Revealed and stated preferences
- Both have pros and cons
- RP: real behavior
- SP: control of the experiment
- It is common to combine them