Choice data

Michel Bierlaire

michel.bierlaire@epfl.ch

Transport and Mobility Laboratory

Data

- About the decision-maker: socio-economic characteristics
  - Collected in any survey
  - Not specific to choice models
  - Collect those that seem relevant for the analysis
- About the alternatives: utility and attributes
  - Utility is an abstract concept
  - Cannot be observed
  - Choice data versus preference data

Preference data

- Consider the following beers:
  1. Cardinal
  2. Kronenbourg
  3. Orval
  4. Tsing Tao
- Method of rating: Associate a rate from 0 (worst) to 10 (best) with each beer
- Method of ranking: Rank the beers, from the best to the worst

Preference data

Drawbacks of rating:

- 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 (cf. Amazon.com)

Preference data

Drawbacks of ranking:

- In general, easy to identify best and worst
- Rank of intermediary alternatives less obvious
- How does the analyst distinguish between real preference and random order?

We need choice data and not preference data

Choice data

- Revealed Preferences (RP)
  - actual choice observed
  - in real market situations
  - Example: scanner data in supermarkets
- Stated Preferences (SP)
  - hypothetical situations
  - attributes defined by the analyst

Example of SP data

- Analysis of response to traffic information in Switzerland
- Project sponsored by Swiss Federal Road Office OFROU
- Analysis by ETHZ and EPFL
- Socio-economic characteristics - choice context
- Stated preference

RP data: advantages

- Real life choices
- Possibility to replicate market shares
- Decision-makers have to assume their choice
- “A bike or a Ferrari?” — “A Ferrari, of course!”
- Real constraints involved

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
- Hard to replicate 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: road pricing)
- Justification bias (example: choice of TV programs)
- 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*

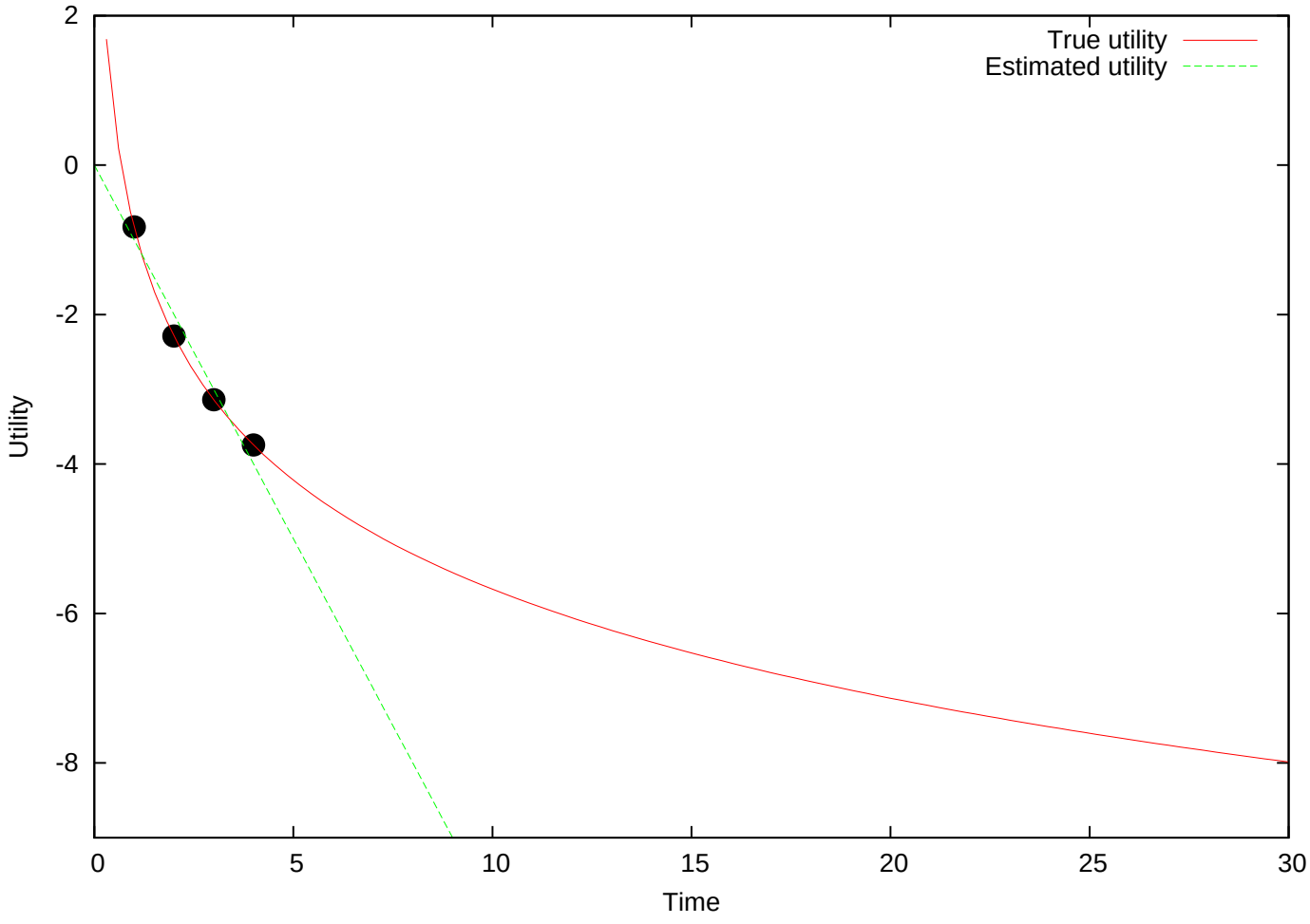
Experimental design

- 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.”

Generation of the 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

Full factorial design

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 :	✔	❒

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Generation of the design

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

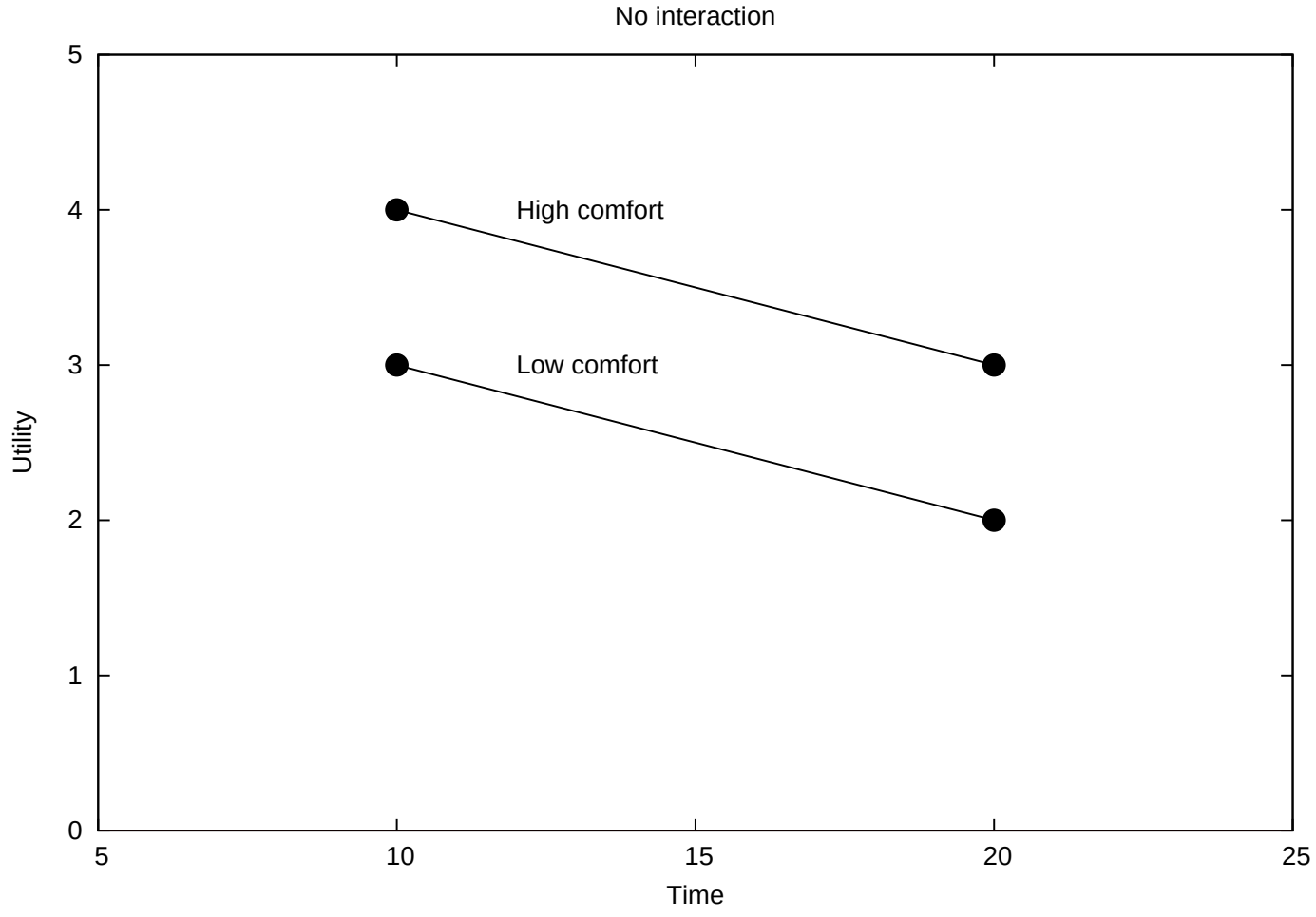
Effects

Main effect

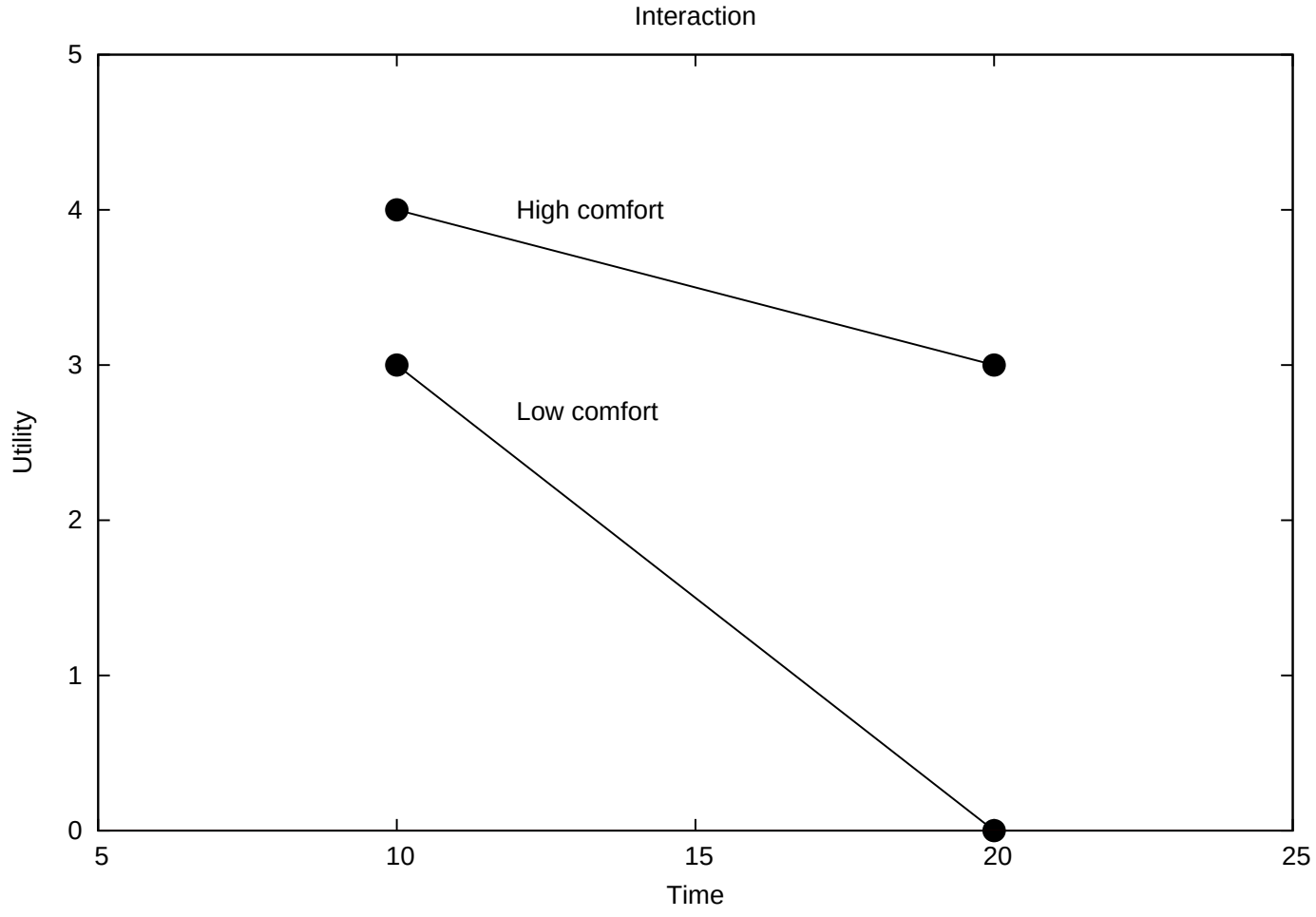
The main effect of a variable is the effect of the experimental response of going from one level of the variable to the next given that the remaining variables do not change

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

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

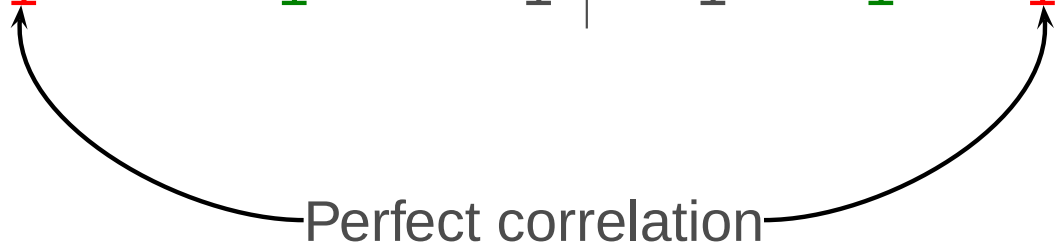
Main effects and interactions

	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



Impossible to distinguish between C-T and model.

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	

Choice data

Each observation must contain

- The socio-economic characteristics of the decision-maker
- For each alternative, the associated attributes
- The choice (revealed or stated)

Combining RP and SP data

- Example: mode choice between car and train (Nijmegen, The Netherlands, 1987)
- Both revealed and stated preference data were collected
- First, use only RP data
- Specification table:

Coefficient	Car	Rail
β^{RP}_{rail}	0	1
β^{RP}_{cost}	Cost	Cost
$\beta^{RP}_{ivttCar}$	in-veh. time	0
$\beta^{RP}_{ovttCar}$	walk time	0
$\beta^{RP}_{ivttRail}$	0	in-veh. time
$\beta^{RP}_{ovttRail}$	0	out-veh. time

Combining RP and SP data

Number of observations = 228

$$\mathcal{L}(0) = -158.038$$
$$\mathcal{L}(\hat{\beta}) = -110.940$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 94.196$$
$$\bar{\rho}^2 = 0.260$$

Variable number		Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	β^{RP}_{rail}	-2.50	1.06	-2.35	0.02
2	$\beta^{RP}_{ivttCar}$	-2.12	0.548	-3.87	0.00
3	$\beta^{RP}_{ivttRail}$	0.380	0.492	0.77	0.44
4	$\beta^{RP}_{ovttCar}$	-1.78	1.59	-1.12	0.26
5	$\beta^{RP}_{ovttRail}$	-2.75	0.889	-3.09	0.00
6	β^{RP}_{cost}	-0.122	0.0263	-4.63	0.00

Combining RP and SP data

- Then, use SP data only
- Specification table:

Coefficient	Car	Rail
β^{SP}_{rail}	0	1
β^{SP}_{inert}	0	1(RP choice is rail)
β^{SP}_{cost}	Cost	Cost
$\beta^{SP}_{ivttCar}$	in-veh. time	0
$\beta^{SP}_{ivttRail}$	0	in-veh. time

Combining RP and SP data

Number of observations = 1511

$$\mathcal{L}(0) = -1047.345$$
$$\mathcal{L}(\hat{\beta}) = -652.193$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 790.305$$
$$\bar{\rho}^2 = 0.373$$

Variable number		Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	β^{SP}_{rail}	-3.13	0.419	-7.47	0.00
2	β^{SP}_{inert}	2.58	0.163	15.78	0.00
3	$\beta^{RP}_{ivttCar}$	-1.02	0.184	-5.52	0.00
4	$\beta^{RP}_{ivttRail}$	-0.266	0.162	-1.64	0.10
5	β^{SP}_{cost}	-0.0389	0.00855	-4.56	0.00

Combining RP and SP data

Compare the coefficients:

Coeff.	RP		SP	
name	est.	t-test	est.	t-test
β_{rail}	-2.50	-2.35	-3.13	-7.47
β_{ivttCar}	-2.12	-3.87	-1.02	-5.52
β_{ivttRail}	0.380	0.77	-0.266	-1.64
β_{ovttCar}	-1.78	-1.12		
β_{ovttRail}	-2.75	-3.09		
β_{cost}	-0.122	-4.63	-0.0389	-4.56
$\beta^{\text{SP}}_{\text{inert}}$			2.58	15.78

Why are the coefficients so different?

How to combine both sets of data into one model?

Combining RP and SP data

- Why are the coefficients so different?
- RP model: $U_{\text{RP}} = V_{\text{RP}} + \varepsilon_{\text{RP}}$
- Normalization of the error term

$$\alpha_{\text{RP}} U_{\text{RP}} = \alpha_{\text{RP}} V_{\text{RP}} + \alpha_{\text{RP}} \varepsilon_{\text{RP}}$$

  where α_{RP} is such that $\text{Var}(\alpha_{\text{RP}} \varepsilon_{\text{RP}}) = \pi^2/6$
- SP model: $U_{\text{SP}} = V_{\text{SP}} + \varepsilon_{\text{SP}}$
- Normalization of the error term

$$\alpha_{\text{SP}} U_{\text{SP}} = \alpha_{\text{SP}} V_{\text{SP}} + \alpha_{\text{SP}} \varepsilon_{\text{SP}}$$

  where α_{SP} is such that $\text{Var}(\alpha_{\text{SP}} \varepsilon_{\text{SP}}) = \pi^2/6$

Combining RP and SP data

We have

$$\frac{\pi^2}{6} = \alpha_{\text{RP}}^2 \operatorname{Var}(\varepsilon_{\text{RP}}) = \alpha_{\text{SP}}^2 \operatorname{Var}(\varepsilon_{\text{SP}})$$

or

$$\frac{\alpha_{\text{RP}}^2}{\alpha_{\text{SP}}^2} = \frac{\operatorname{Var}(\varepsilon_{\text{SP}})}{\operatorname{Var}(\varepsilon_{\text{RP}})}$$

- There is no reason to have $\operatorname{Var}(\varepsilon_{\text{SP}}) = \operatorname{Var}(\varepsilon_{\text{RP}})$
- Therefore, there is no reason to have $\alpha_{\text{SP}}^2 = \alpha_{\text{RP}}^2$
- For a given β, what is estimated is
  - $\alpha_{\text{RP}}\beta$ in the RP model
  - $\alpha_{\text{SP}}\beta$ in the SP model

Combining RP and SP data

How to combine both sets of data into one model?

- Define a model with 4 alternatives.

Coefficient	CarRP	RailRP	CarSP	Rail SP
β^{RP}_{rail}	0	1	0	0
$\beta^{RP}_{ovttCar}$	walk time	0	0	0
$\beta^{RP}_{ovttRail}$	0	out-veh. time	0	0
$\beta^{RPSP}_{ivttCar}$	in-veh. time	0	in-veh. time	0
$\beta^{RPSP}_{ivttRail}$	0	in-veh. time	0	in-veh. time
β^{RPSP}_{cost}	Cost	Cost	Cost	Cost
β^{SP}_{rail}	0	0	0	1
β^{SP}_{inert}	0	0	0	1(RP choice is rail)

Combining RP and SP data

- For RP observations, declare the 2 SP alt. unavailable
- For SP observations, declare the 2 RP alt. unavailable
- So, each observation corresponds to 2 alternatives.
- Explicitly include the scales

$$\begin{aligned} V_{\text{RP}} &= \alpha_{\text{RP}}\beta^{\text{RPSP}}x_1 + \alpha_{\text{RP}}\beta^{\text{RP}}x_2 \\ V_{\text{SP}} &= \alpha_{\text{SP}}\beta^{\text{RPSP}}x_1 + \alpha_{\text{SP}}\beta^{\text{SP}}x_3 \end{aligned}$$

- Normalize the RP scale to 1 and estimate the SP scale

$$\begin{aligned} V_{\text{RP}} &= \beta^{\text{RPSP}}x_1 + \beta^{\text{RP}}x_2 \\ V_{\text{SP}} &= \alpha_{\text{SP}}\beta^{\text{RPSP}}x_1 + \alpha_{\text{SP}}\beta^{\text{SP}}x_3 \end{aligned}$$

Combining RP and SP data

Number of observations = 1739

$$\begin{aligned} \mathcal{L}(0) &= -1205.383 \\ \mathcal{L}(\hat{\beta}) &= -764.613 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 881.541 \\ \bar{\rho}^2 &= 0.358 \end{aligned}$$

Combining RP and SP data

Variable number		Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	β^{RP}_{rail}	-1.99	0.921	-2.16	0.03
2	β^{SP}_{rail}	-8.62	2.01	-4.28	0.00
3	β^{SP}_{inert}	6.59	1.60	4.12	0.00
4	$\beta^{RPSP}_{ivttCar}$	-2.43	0.471	-5.16	0.00
5	$\beta^{RPSP}_{ivttRail}$	-0.244	0.331	-0.74	0.46
6	$\beta^{RP}_{ovttCar}$	-1.42	1.62	-0.88	0.38
7	$\beta^{RP}_{ovttRail}$	-2.88	0.888	-3.25	0.00
8	β^{RPSP}_{cost}	-0.110	0.0240	-4.61	0.00
9	α_{SP}	0.382	0.0921	-6.72[a]	0.00

[a] t-test against 1

Combining RP and SP data

Remember:

$$\frac{\alpha_{\mathsf{RP}}^2}{\alpha_{\mathsf{SP}}^2} = \frac{\mathrm{Var}(\varepsilon_{\mathsf{SP}})}{\mathrm{Var}(\varepsilon_{\mathsf{RP}})}$$

Here:

$$\frac{1}{0.382} = \frac{\mathrm{Var}(\varepsilon_{\mathsf{SP}})}{\mathrm{Var}(\varepsilon_{\mathsf{RP}})} = 2.62 > 1$$

Therefore

$$\mathrm{Var}(\varepsilon_{\mathsf{SP}}) > \mathrm{Var}(\varepsilon_{\mathsf{RP}})$$

which is consistent with intuition.

Combining RP and SP data

Warning: the utility fonction is not linear-in-parameter anymore

$$
\begin{aligned}
V_{\text{RP}} &= \beta^{\text{RPSP}} x_1 + \beta^{\text{RP}} x_2 \\
V_{\text{SP}} &= \alpha_{\text{SP}} \beta^{\text{RPSP}} x_1 + \alpha_{\text{SP}} \beta^{\text{SP}} x_3
\end{aligned}
$$

Warning:

- unobserved individual heterogeneity is ignored in this framework
- mixtures of models are required to capture them