

# Discrete Choice Models and Related Issues

Stephane Hess

# Outline



UNIVERSITY OF LEEDS

- What do we do with SP data?
- Software requirements
- Estimation example
- Available packages
- Introduction to Biogeme
- An example of unexpected results
- A word on joint RP/SP estimation

# SP data uses



UNIVERSITY OF LEEDS

- SP data is primarily useful to gauge relative sensitivities
- Main example: WTP measures
- Use on its own in forecasting or elasticity computation is not advisable (though sometimes done)
  - absolute response may be very different from real life sensitivities
- Can jointly estimate models on SP and RP data
  - SP data provides relative sensitivities, RP data absolute sensitivities (scale)

# Software requirements



UNIVERSITY OF LEEDS

- Correct results
  - goes without saying
- Stability
- Flexibility
  - Model types
  - Utility functions
- Speed
- Output
- User friendliness



# A word on estimation...

- Need to find parameter values that reproduce observed choices
- Have log-likelihood (LL) function, conditional on  $\beta$

$$LL(\beta) = \sum_{n=1}^N \ln(P_{n,j_n}(\beta))$$

- Need to maximise LL in relation to  $\beta$ , find maximum LL estimate of  $\beta$
- At MLE, we have:

$$\frac{\partial LL(\beta)}{\partial \beta} = 0$$

# Estimation of MNL I



UNIVERSITY OF LEEDS

- MNL choice probabilities have closed form expression
- No simulation required in estimation
- Log-likelihood for MNL with linear in parameters utility specification is globally concave
  - if a solution exists, it is unique



# Estimation of MNL II

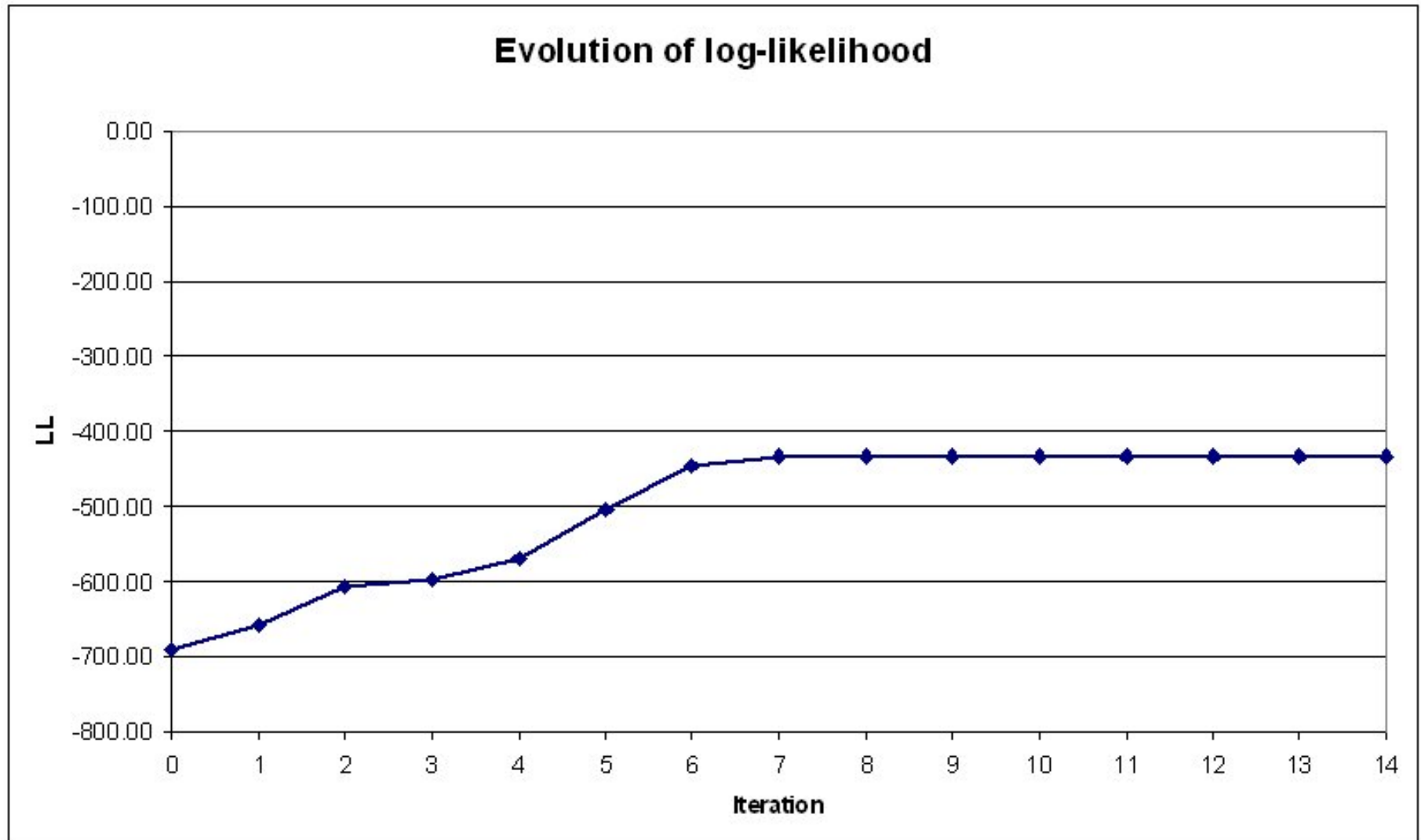
- Estimation example
- Swissmetro data
  - three alternatives: car, train, swissmetro

- Basic utility specification

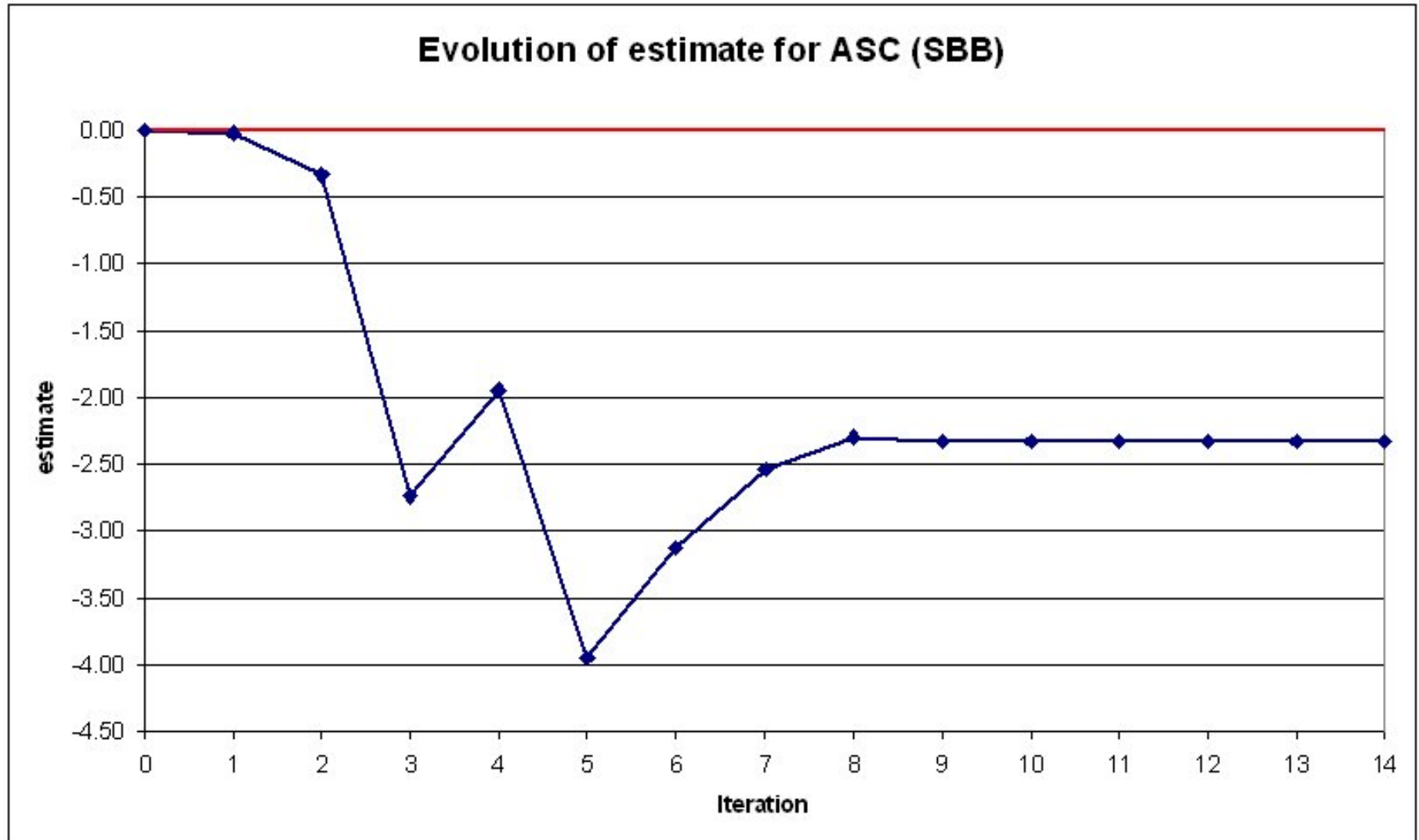
$$V_i = \delta_i + \beta_{TT}TT_i + \beta_{TC}TT_i$$

- Alternative specific constant ( $\delta_i$ ) set to zero for car
- Maximum likelihood estimation searches for  $\beta$  and  $\delta$  values that maximise likelihood of observed choices

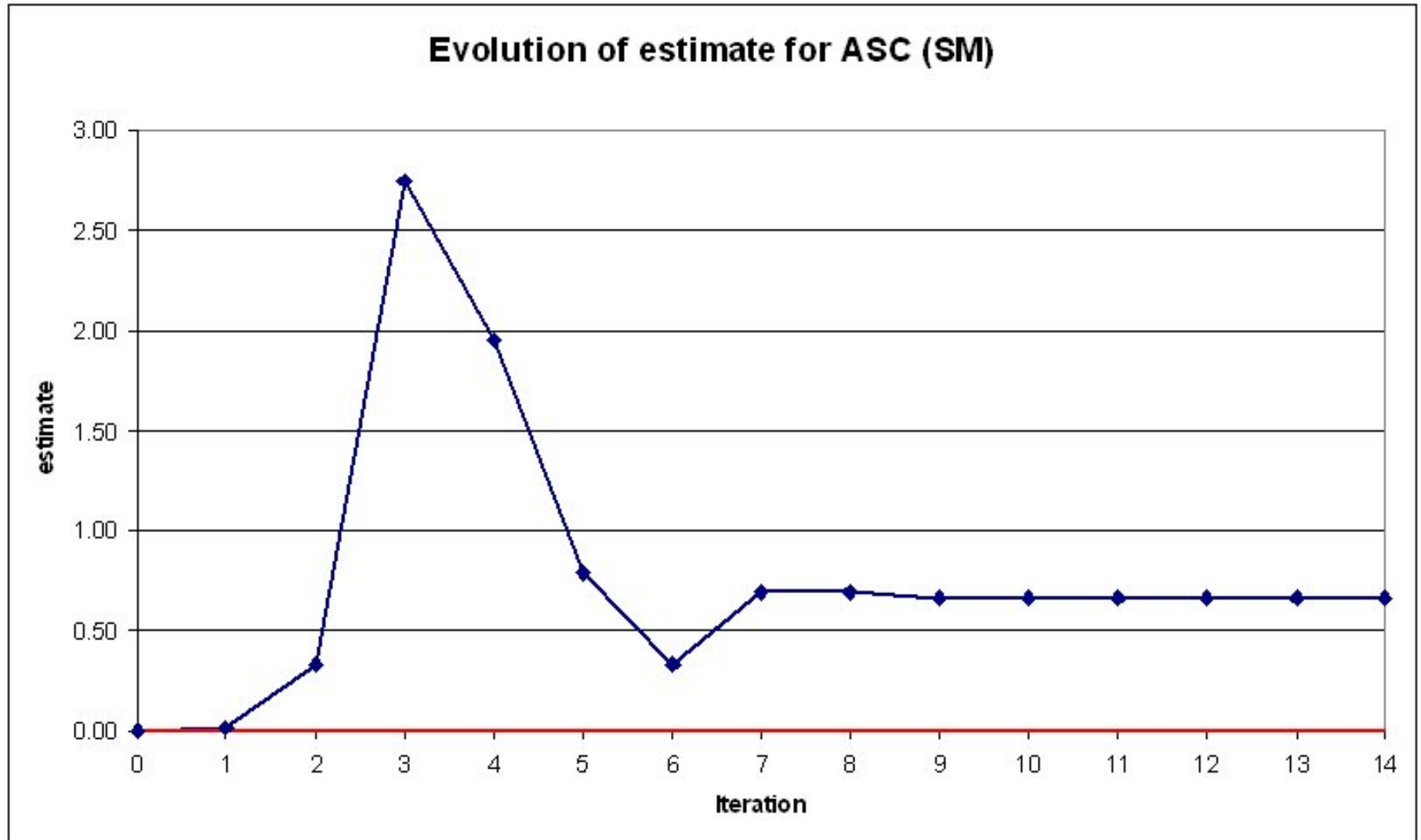
# Estimation of MNL III



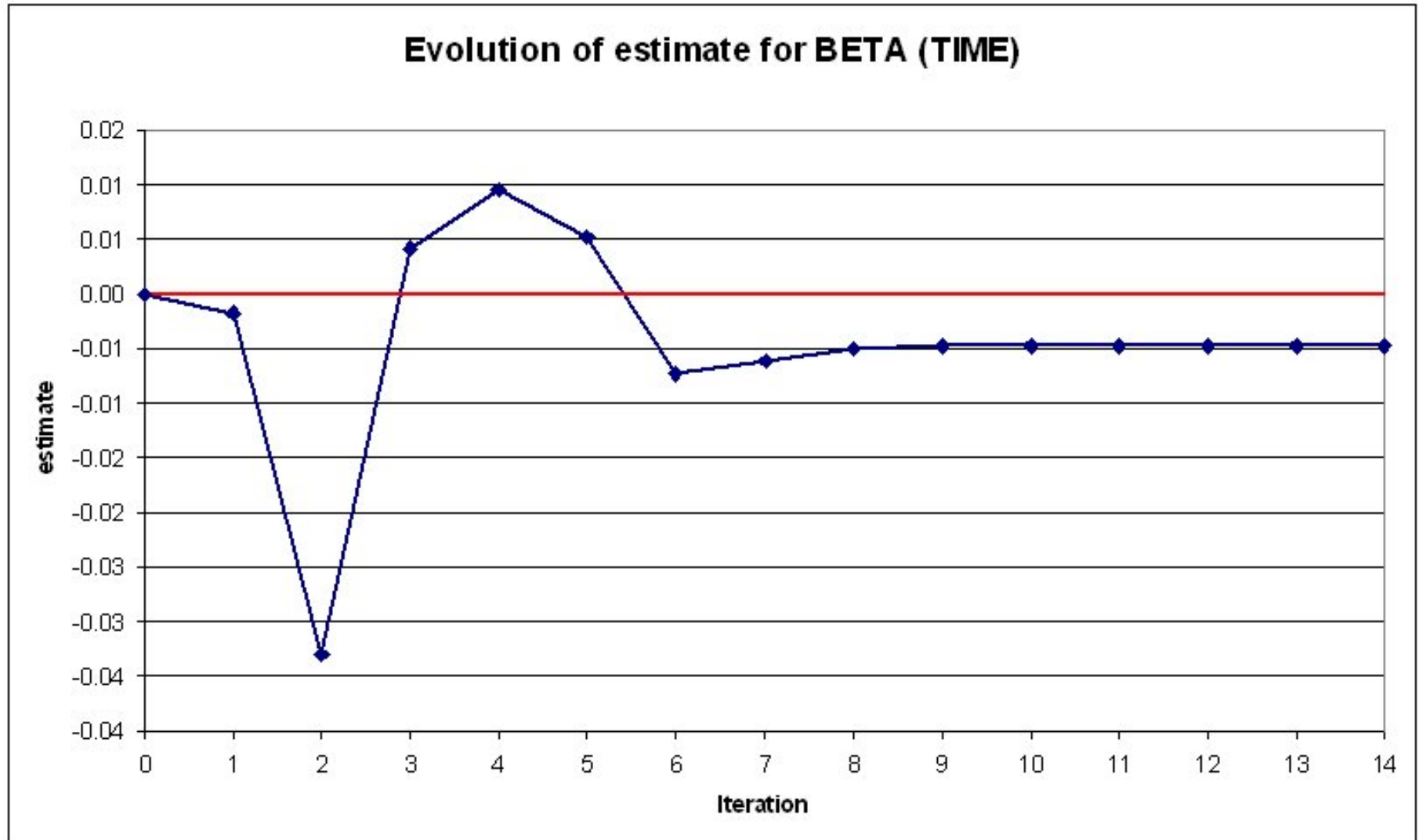
# Estimation of MNL IV



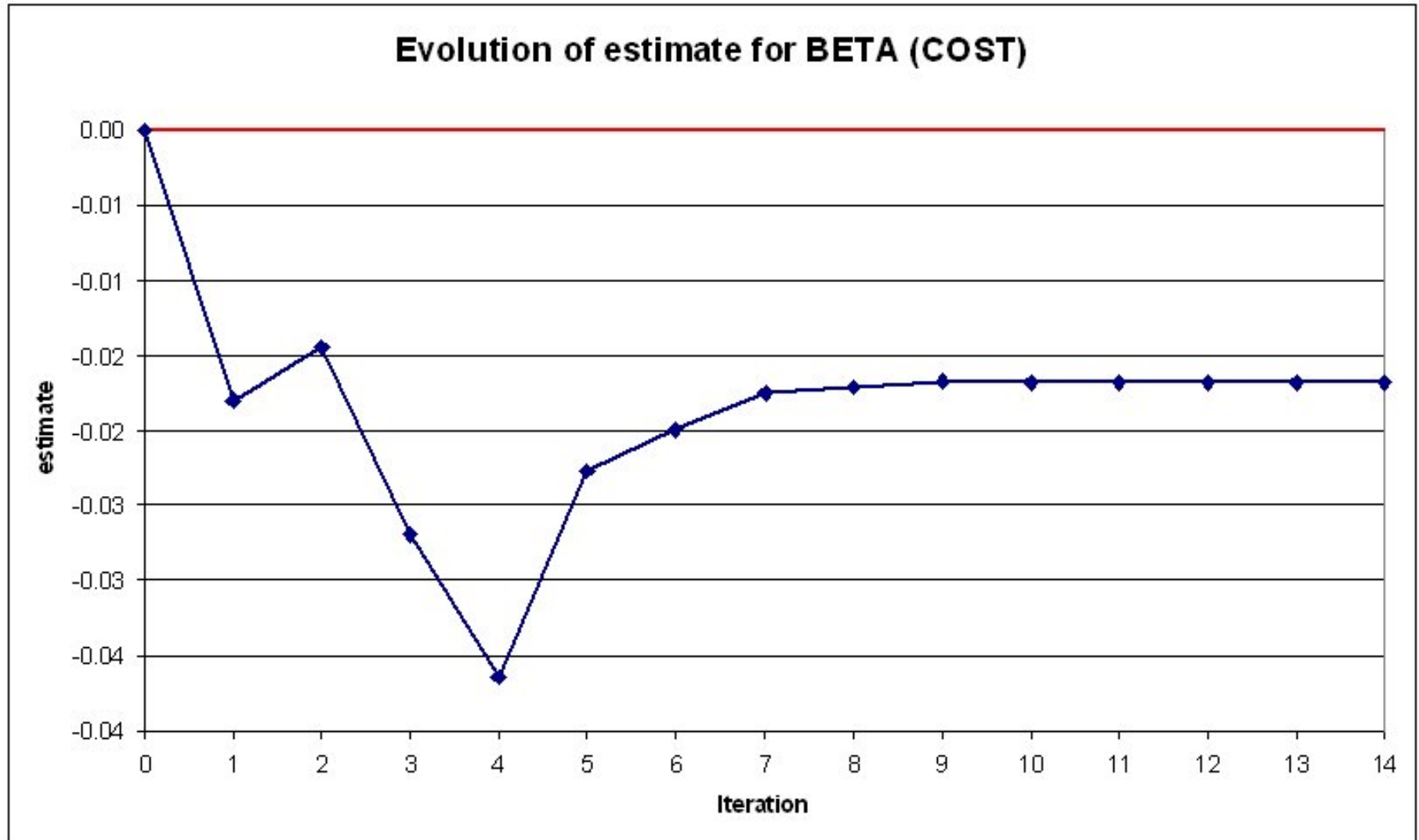
# Estimation of MNL V



# Estimation of MNL VI



# Estimation of MNL VII



# Estimation I



UNIVERSITY OF LEEDS

- Basic example
  - Two alternatives
  - Two attributes
    - \* Travel time (TT) in minutes
    - \* Toll in £

# Estimation II

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2														
3					beta(TT)	beta(TOLL)		VTTS (GBP/hr)						
4					-0.1	-0.5		12						
5														
6				Alternative	TT (min)	TOLL (GBP)		V(i)						P(i)
7				1	30	0		-3		exp[V(2)-V(1)]	1			50.00%
8				2	25	1		-3		exp[V(1)-V(2)]	1			50.00%
9														
10														

- Attributes of alternative 1 in cells E7 and F7
- Attributes of alternative 2 in cells E8 and F8
- Marginal utility coefficients in cells E4 and E5
- Calculated VTTS in H4, willingness to pay for red. in TT by 1 hr.
- Observed utilities (calculated) in column H
- Exponential of differences in utilities (calculated) in column K
- Choice probabilities (calculated) in column N

- Current choice probabilities
  - Moving from alternative 1 to alternative 2 saves 5 minutes but costs £1
  - Equates to £12 for 1 hour
    - \* equal to VTTS with current coefficients
    - \* respondents indifferent between two alternatives
- Excel example ...
  - How do choice probabilities change if we change relative coefficient values (and hence VTTS)?
  - How do choice probabilities change if we change absolute coefficient values (scale effects)?

# Estimation IV



- Actual estimation example
- Have 100 observations, with observed choices
- Same alternative structure as first example

	ABC	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																	
2																	
3				beta(TT)	beta(TOLL)		VTTs (GBP/hr)										
4				0	0		#DIV/0!										
5																	
6				Alternative 1		Alternative 2			V(i)		exp[V(diff)]		P(i)				
7		Observation	TT (min)	TOLL (GBP)	TT (min)	TOLL (GBP)	Choice	Alt 1	Alt 2	exp[V(2)-V(1)]	exp[V(1)-V(2)]	Alt 1	Alt 2			Contribution to LL	
8		1	26	4	35	1	2	0	0	1	1	50.00%	50.00%			-0.693147181	
9		2	27	3	33	1	2	0	0	1	1	50.00%	50.00%			-0.693147181	
10		3	22	3	20	6	1	0	0	1	1	50.00%	50.00%			-0.693147181	
11		4	24	5	37	0	2	0	0	1	1	50.00%	50.00%			-0.693147181	
12		5	25	4	30	3	1	0	0	1	1	50.00%	50.00%			-0.693147181	
13		6	29	4	30	3	1	0	0	1	1	50.00%	50.00%			-0.693147181	
14		7	29	3	22	3	2	0	0	1	1	50.00%	50.00%			-0.693147181	
15		8	34	0	28	2	1	0	0	1	1	50.00%	50.00%			-0.693147181	
16		9	23	4	31	3	1	0	0	1	1	50.00%	50.00%			-0.693147181	
17		10	22	5	24	5	2	0	0	1	1	50.00%	50.00%			-0.693147181	
18		11	34	1	35	2	2	0	0	1	1	50.00%	50.00%			-0.693147181	
19		12	20	4	38	0	2	0	0	1	1	50.00%	50.00%			-0.693147181	
20		13	39	0	31	3	2	0	0	1	1	50.00%	50.00%			-0.693147181	
21		14	32	2	31	4	1	0	0	1	1	50.00%	50.00%			-0.693147181	
22		15	22	4	28	4	1	0	0	1	1	50.00%	50.00%			-0.693147181	
23		16	34	0	36	1	1	0	0	1	1	50.00%	50.00%			-0.693147181	
24		17	22	5	32	3	2	0	0	1	1	50.00%	50.00%			-0.693147181	
25		18	29	2	30	1	2	0	0	1	1	50.00%	50.00%			-0.693147181	
26		19	38	2	25	2	2	0	0	1	1	50.00%	50.00%			-0.693147181	

# Estimation V



- Spreadsheet calculates contribution to log-likelihood (LL)
  - logarithm of choice probability of chosen alternative
- Total LL given by sum of individual contributions
- Task:
  - Currently marginal utility coefficients set to zero
  - Choice probabilities equal across two alternatives
    - \* all observed utilities equal to zero
    - \* all behaviour explained by error
  - See how LL evolves as we change values of marginal utility coefficients (cells E4 and F4)
  - Try using Solver to find optimal values

- Inputs for estimation packages
  - data with observed (or stated) choices
  - model structure (e.g. binary logit, MNL, ...)
  - specification of utility function
  - estimation settings
- Estimation packages maximise likelihood of observed (stated) choices by changing values of model parameters
- Outputs from estimation packages
  - estimates of model parameters
  - model fit statistics, and possibly diagnostic statistics
  - potentially trade-offs between coefficients (easy to calculate)
  - potentially elasticities and forecasts

# Available packages I



UNIVERSITY OF LEEDS

- Number of different packages available
- Advantages and disadvantages
- Main trade-offs involved:
  - Freeware vs. commercial
  - Speed vs. flexibility

# Available packages II



UNIVERSITY OF LEEDS

- **ALogit**

- commercial
- in use for over 25 years
- very significant speed advantages
- advantages in application
- can be used for MNL, NL and MMNL
- restrictions on utility specifications

# Available packages III



UNIVERSITY OF LEEDS

- **BIOGEME**

- freeware
- since 2003
- great flexibility
  - \* model structures
  - \* utility specification
- some issues with speed

# Available packages IV



UNIVERSITY OF LEEDS

- **Limdep/NLogit**
  - commercial
  - speed advantages
  - data analysis features
  - many different model forms

# Available packages V



UNIVERSITY OF LEEDS

- Other packages
  - Kenneth Train's Gauss code
  - OxDCM
  - LatentGold
  - Ngene
  - others...

# Introduction to Biogeme



UNIVERSITY OF LEEDS

- Biogeme requires three files
  - model file (.mod)
  - data file (tab separated, with variable names in first row)
  - parameter file (.par)

# Biogeme: data file



UNIVERSITY OF LEEDS

- **Data file**
- Contents:
  - choice variable (required)
  - attributes of alternatives
  - attributes of decision makers
- Format:
  - Header row with names of variables
  - One row for each observation (choice)

# Biogeme: model file I

- **Model file**
- Contains all information on model structure and utility function
- Split into different sections, all defined as

```
[Section]
```

- Need spaces between individual statements!
- Define choice variable

```
[Choice]
```

```
CHOICE
```

# Biogeme: model file II

- Define taste coefficients
  - for each coefficient, have:
    - \* name
    - \* starting value
    - \* lower bound
    - \* upper bound
    - \* status (estimated = 0 or fixed = 1)

```
[Beta]
// Name Value LowerBound UpperBound status
ASC_CAR 0 -10 10 0
BETA_TT 0 -10 10 0
BETA_TC 0 -10 10 0
```

# Biogeme: model file III



UNIVERSITY OF LEEDS

- Utility functions
  - “Id” refers to code used in choice variable
  - “Name” only for output
  - “Avail” gives variable defining availabilities (0/1)

```
[Utilities]
// Id Name Avail linear-in-parameter expression (beta1*x1 + beta2*x2 + ... )
1 SBB TRAIN_AV BETA_TT * TRAIN_TT + BETA_TC * TRAIN_CO
2 SM SM_AV ASC_SM * one + BETA_TT * SM_TT + BETA_TC * SM_CO
3 Car CAR_AV ASC_CAR * one + BETA_TT * CAR_TT + BETA_TC * CAR_CO
```

- Any non-linear statements need to go in [GeneralizedUtilities]

# Biogeme: model file IV

- Expressions section
  - defines variables not included in data file

```
[Expressions]  
one = 1
```

- Exclude section
  - exclude some observations from analysis
  - here: exclude all non-commuters

```
[Exclude]  
PURPOSE != 1
```

# Biogeme: model file V



UNIVERSITY OF LEEDS

- Model section

```
[Model]
$MNL // Multinomial Logit Model
```

- Produce ratios of coefficients

```
[Ratios]
// Numerator Denominator Name
beta_tt beta_tc vtts
```

# Biogeme: model file VI



UNIVERSITY OF LEEDS

- Some advanced specifications
- Weight observations differently

```
[Weight]
```

- Recognise repeated choice nature of data (MMNL only)

```
[PanelData]
```

# Biogeme: model file VII

- Random distribution of MMNL models (e.g. with Normal)

– replace

```
BETA_TT * TRAIN_TT
```

by

```
BETA_TT [ beta_tt_sig ] * TRAIN_TT
```

- Number of draws to be used in MMNL simulation

```
[Draws]
```

- Divide population into groups

```
[Group]
```

# Biogeme: model file VIII

- Different scale in different groups

```
[Scale]
// Group_number  scale LowerBound UpperBound status
1 1 0.1 10 1
2 1 0.1 10 0
```

- Other model forms:

```
$BP // Binary Probit Model
$OL // Ordinal logit
$NL // Nested Logit Model
$CNL // Cross-Nested Logit Model
$NGEV // Network GEV Model
```

# Biogeme: parameter file



UNIVERSITY OF LEEDS

- either generic file (“default.par”) or file specific to model (“model\_1.par”)
- most basic specification

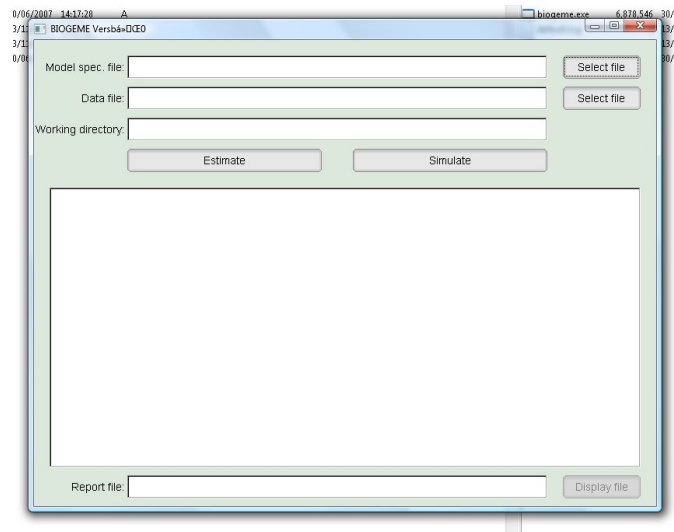
```
[GEV]
gevAlgo = "BIO"
gevScreenPrintLevel = 1
gevLogFilePrintLevel = 2
```

- defines estimation algorithm and level of detail for screen and file output

# Biogeme: execution I

- Call to Biogeme via DOS command prompt
  - “biogeme model\_1 swissmetro.dat”
- Use Windows shell

`winbiogeme.exe`



# Biogeme: execution II



UNIVERSITY OF LEEDS

- For those with limited DOS knowledge, and avoiding problems with Windows version...
- Batch file for running Biogeme: run\_biogeme.bat
- Before running first time, need to edit batch file and specify location of biogeme.exe
- Then copy batch file in directory with model and data file and launch

# Biogeme: execution III



```
15/2009 15:14:33 A reference library
15/2009 C:\WINDOWS\system32\cmd.exe
15/2009 Enter the name of your model file, without the .mod extension: model
15/2009 Enter the name of your data file, including the extension: data_new.txt
15/2009 -----
15/2009 Launching Biogeme...
15/2009 -----
15/2009 BIOGEME Version 1.6 [Sun Mar 16 16:06:18 GMT 2008]
15/2009 Michel Bierlaire, EPFL
15/2009 -- Compiled by Michel on MINGW32_NT-5.1
15/2009 See http://biogeme.epfl.ch
15/2009 !! CFSQP is available !!
15/2009 -----
15/2009 "In every non-trivial program there is at least one bug."
15/2009
15/2009 [15:16:04]patFileNames.cc:68 model.par does not exist
15/2009 [15:16:04]patFileNames.cc:72 Trying default.par instead
15/2009 [15:16:04]patBiogeme.cc:131 Read default.par
15/2009 Warning: Lower bound on mu set to 1e-005
13/2009 Warning: Value defined by gevMinimumMu in default.par
13/2009 Opening file data_new.txt
15/2009 Data file... line 500 Memory: 80 Kb
15/2009 Total obs.: 899
15/2009 Total memory: 145.898 Kb
15/2009 Run time for data processing: 00:00
```

# Biogeme: outputs



UNIVERSITY OF LEEDS

- Output files
  - log file: output at each iteration
  - html and rep file: model results
  - res file: same format as mod file, but with final estimates used for coefficients
  - sta file: data statistics
  - tex file: outputs in LaTeX format
  - summary.html: summary file for all models, using parameters defined in .lis file

# Unexpected results I



UNIVERSITY OF LEEDS

- Swissmetro data
- Three alternatives
  - Train (SBB)
  - Swissmetro (SM)
  - Car
- Attributes
  - Travel time (all alternatives)
  - Travel cost (all alternatives)
  - Headway (SBB and SM only)
  - Airline style seats (SM only)

# Unexpected results II



- Base model results

Observations	1575	
Final LL	-1180.88	
Par.	6	
	est.	t-rat.
ASC_CAR	0.311	1.72
ASC_SM	1.24	7.09
BETA_HW	-0.00697	-2.99
BETA_TC	0.000173	4.72
BETA_TT	-0.00345	-1.98
DELTA_AL_SEAT	0.0581	0.37

# Unexpected results III

- Take into account season tickets

Observations	1575	
Final LL	-1120.9	
Par.	6	
	est.	t-rat.
ASC_CAR	0.177	0.98
ASC_SM	1.44	8.52
BETA_HW	-0.00705	-3.05
BETA_TC	-0.0104	-8.73
BETA_TT	-0.00319	-2.02
DELTA_AL_SEAT	0.0727	0.44



# Joint estimation

- Especially useful if combining RP and SP data with a view to correcting scale differences
- Also useful when combining e.g. mode choice and route choice data
- Relative sensitivities are kept constant across datasets, absolute sensitivities are allowed to vary
- Set scale in a base group to 1
- Estimates for marginal utility coefficients are given at response level for that base group, e.g. the RP data

# NL approach I



UNIVERSITY OF LEEDS

- Group alternatives from different data sources into individual nests
  - sometimes referred to as *NL-trick* as not a real nesting structure
  - only alternatives from one nest available at any given time
- Structural parameters linked to scale and hence variance of error term
- Different structural parameters in different nests

# NL approach II



UNIVERSITY OF LEEDS

- Impose constraint on one of the structural parameters (relative values)
- Structural parameters no longer necessarily lower than 1
  - different role to traditional NL
- Issues:
  - different NL implementations
  - gets complicated if we additionally want to allow for random heterogeneity, or for inter-alternative correlation

# Direct rescaling approach I



UNIVERSITY OF LEEDS

- Remember:
  - only differences in utility matter
  - need to normalise location and scale
- Typical normalisation for MNL:
  - set  $\mu = 1$  and  $\eta = 0$
  - variance becomes  $\frac{\pi^2}{6} \sim 1.6449$
- $P_n(i) = \frac{e^{V_{n,i}}}{\sum_{j=1}^J e^{V_{n,j}}}$
- with scale parameter:  $P_n(i) = \frac{e^{\mu V_{n,i}}}{\sum_{j=1}^J e^{\mu V_{n,j}}}$

# Direct rescaling approach II



UNIVERSITY OF LEEDS

- Limiting cases of MNL

- Remember:  $var(\varepsilon_{n,j}) = \frac{\pi^2}{6\mu^2}$

$$P_{n,i} = \frac{e^{\mu V_{n,i}}}{\sum_{j=1}^J e^{\mu V_{n,j}}}$$

- $\lim_{\mu \rightarrow 0} P_{n,i} = \frac{1}{J}$
- $\lim_{\mu \rightarrow \infty} P_{n,i} = 1$  if  $V_{n,i} = \max V_{n,1}, \dots, V_{n,J}$
- Increasing  $\mu$  means lower variance for error term
- Larger relative weight for observed part of utility

# Direct rescaling approach III

- Illustration: allow for differences in scale across data sources
- $D$  different groups
- $U_{i,d} = V_{i,d} + \varepsilon_{i,d}$  where  $var(\varepsilon_{i,d}) = \frac{\pi^2}{6\mu_d^2}$
- Use group 1 as base (arbitrary)

$$U_{i,1} = V_{i,1} + \varepsilon_{i,1}$$

...

$$\alpha_d U_{i,d} = \alpha_d V_{i,d} + \alpha_d \varepsilon_{i,d}$$

...

$$\alpha_D U_{i,D} = \alpha_D V_{i,D} + \alpha_D \varepsilon_{i,D}$$

# Direct rescaling approach IV

- For estimation as a homoscedastic model, we thus obtain that

$$\text{var}(\varepsilon_{i,1}) = \alpha_d^2 \text{var}(\varepsilon_{i,d})$$

- then...

$$\alpha_d^2 = \frac{\text{var}(\varepsilon_{i,1})}{\text{var}(\varepsilon_{i,d})}$$

- and

$$\alpha_d = \frac{\mu_d}{\mu_1}$$

- $\alpha_d$  is larger than 1, variance of unobserved utility components in sample  $d$  is smaller than in base sample
- Done automatically using [Scale] section in BIOGEME

# QUESTIONS?