Estimating hybrid choice models with the new version of Biogeme

Michel Bierlaire

`transp-or.epfl.ch`

Transport and Mobility Laboratory, EPFL

Motivation

- Standard random utility assumptions are usually violated
- Factors such as attitudes, perceptions, knowledge are not reflected

Example: perceptions

Exemple: subscription to *The Economist*

Web only	@ \$59
Print only	@ \$125
Print and web	@ \$125

Example: perceptions

Exemple: subscription to *The Economist*

Experiment 1	Experiment 2
Web only @ \$59	Web only @ \$59
Print only @ \$125	
Print and web @ \$125	Print and web @ \$125

Example: perceptions

Exemple: subscription to *The Economist*

	Experiment 1	Experiment 2	
16	Web only @ $59	Web only @ $59	68
0	Print only @ $125		
84	Print and web @ $125	Print and web @ $125	32

Source: Ariely (2008)

- Dominated alternative
- According to utility maximization, should not affect the choice
- But it affects the perception, which affects the choice.

Example: perceptions

Population of 600 is threatened by a disease. Two alternative treatments to combat the disease have been proposed.

Experiment 1 # resp. = 152	Experiment 2 # resp. = 155
Treatment A: 200 people saved	Treatment C: 400 people die
Treatment B: 600 people saved with prob. 1/3 0 people saved with prob. 2/3	Treatment D: 0 people die with prob. 1/3 600 people die with prob. 2/3

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Example: perceptions

Population of 600 is threatened by a disease. Two alternative treatments to combat the disease have been proposed.

	Experiment 1 # resp. = 152	Experiment 2 # resp. = 155	
72%	Treatment A: 200 people saved	Treatment C: 400 people die	22%
28%	Treatment B: 600 people saved with prob. 1/3 0 people saved with prob. 2/3	Treatment D: 0 people die with prob. 1/3 600 people die with prob. 2/3	78%

Source: Tversky & Kahneman (1986)

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

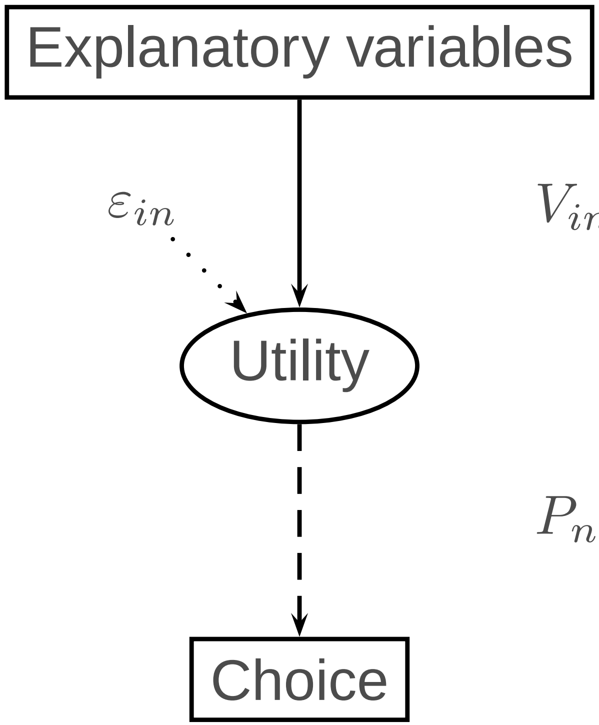
Latent concepts

- **latent**: potentially existing but not presently evident or realized (from old French: hidden)
- Here: not directly observed
- Standard models are already based on a latent concept: utility

Drawing convention:

- Latent variable
- Observed variable
- structural relation: ⟶
- measurement: - - - →
- errors: . . . . . . →

Random utility



$$V_{in} = \sum_k \beta_{ik} x_{ikn}$$

$$P_n(i) = e^{V_{in}} / \sum_j e^{V_{jn}}$$

Attitudes

- Psychometric indicators
- Example: attitude towards the environment.
- For each question, response on a scale: strongly agree, agree, neutral, disagree, strongly disagree, no idea.
  - The price of oil should be increased to reduce congestion and pollution
  - More public transportation is necessary, even if it means additional taxes
  - Ecology is a threat to minorities and small companies.
  - People and employment are more important than the environment.
  - I feel concerned by the global warming.
  - Decisions must be taken to reduce the greenhouse gas emission.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Indicators

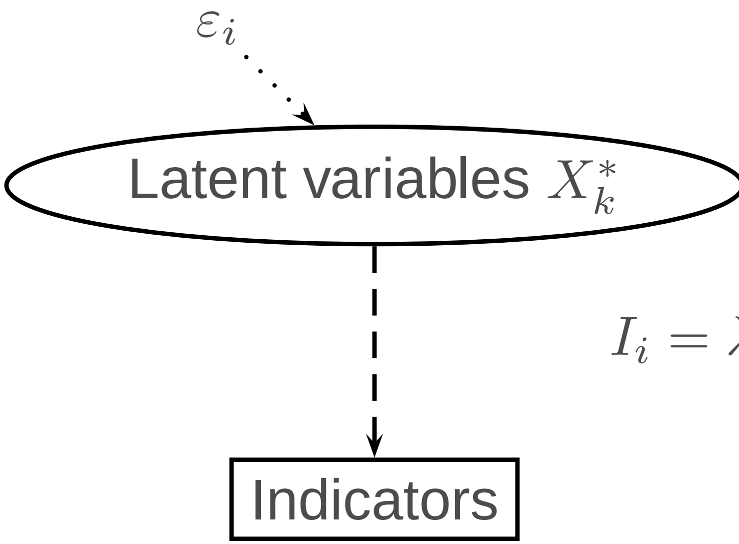
Indicators cannot be used as explanatory variables. Mainly two reasons:

1. Measurement errors
   - Scale is arbitrary and discrete
   - People may overreact
   - Justification bias may produce exaggerated responses
2. No forecasting possibility
   - No way to predict the indicators in the future

TRANSP-OR

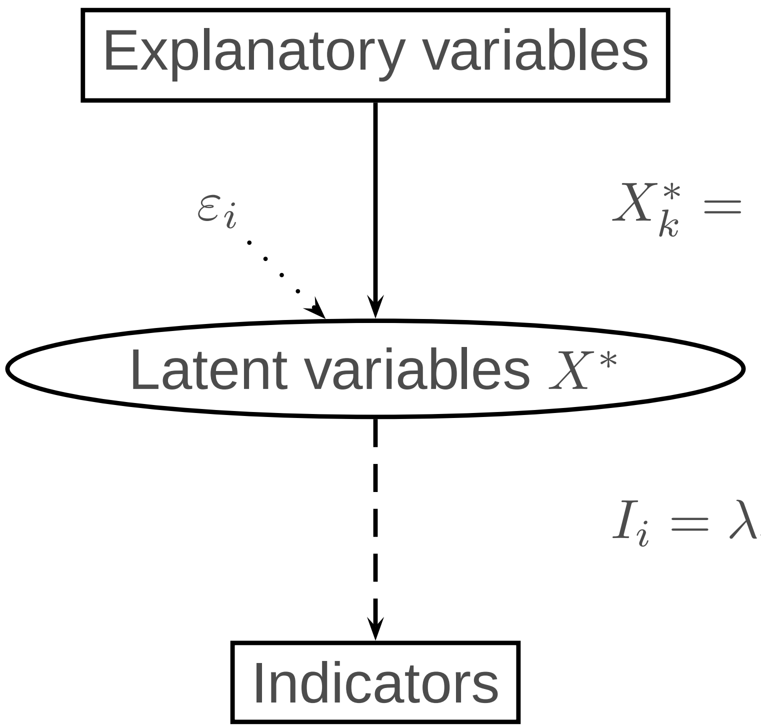
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Factor analysis



$$I_i = \lambda_i + \sum_k L_{ik} X_k^*$$

Measurement equation



$$X_k^* = \sum_j \beta_j x_j$$

$$I_i = \lambda_i + \sum_k L_{ik} X_k^*$$

Measurement equation

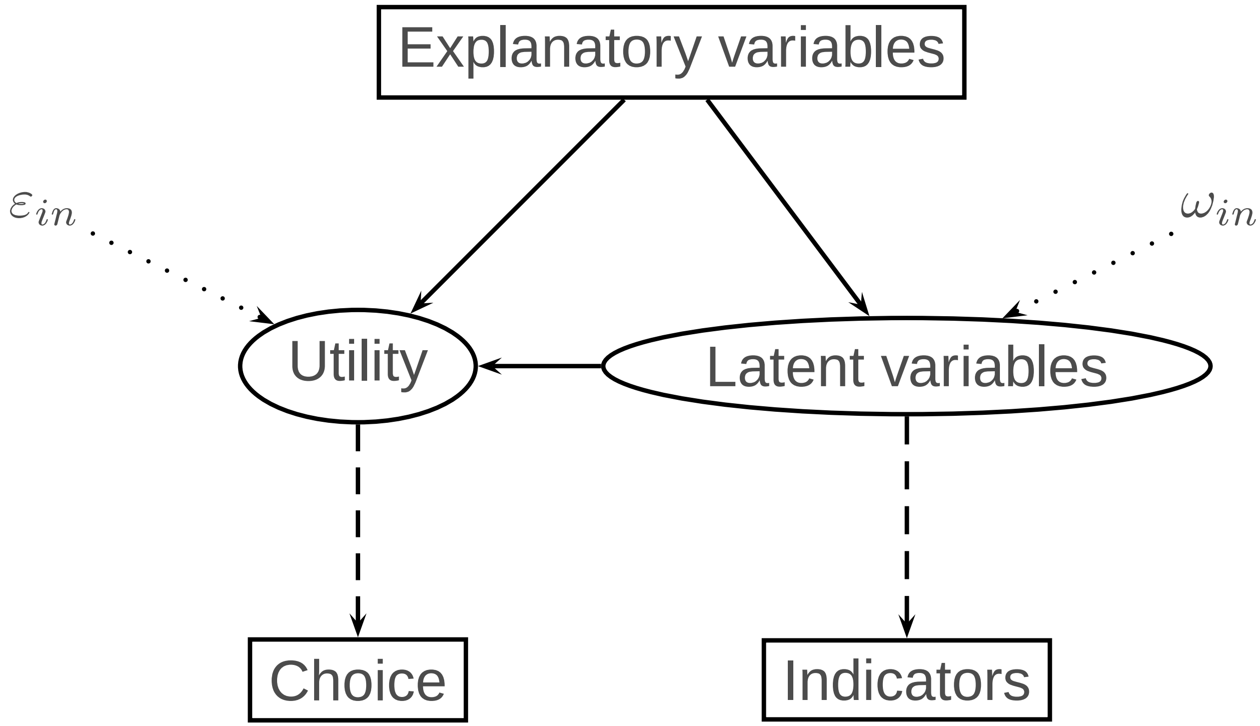
Continuous model: regression

$$I = f(X^*; \beta) + \varepsilon$$

Discrete model: thresholds

$$I = \begin{cases} 1 & \text{if } -\infty < X^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < X^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < X^* \leq \tau_3 \\ 4 & \text{if } \tau_3 < X^* \leq \tau_4 \\ 5 & \text{if } \tau_4 < X^* \leq +\infty \end{cases}$$

Choice model



Estimation: likelihood

Structural equations:

1. Distribution of the latent variables:

$$f_1(X_n^* | X_n; \lambda, \Sigma_\omega)$$

   For instance

$$X_n^* = h(X_n; \lambda) + \omega_n, \quad \omega_n \sim N(0, \Sigma_\omega).$$

2. Distribution of the utilities:

$$f_2(U_n | X_n, X_n^*; \beta, \Sigma_\varepsilon)$$

   For instance

$$U_n = V(X_n, X_n^*; \beta) + \varepsilon_n, \quad \varepsilon_n \sim N(0, \Sigma_\omega).$$

Estimation: likelihood

Measurement equations:

1. Distribution of the indicators:

$$f_3(I_n | X_n, X_n^*; \alpha, \Sigma_\nu)$$

   For instance:

$$I_n = m(X_n, X_n^*; \alpha) + \nu_n, \quad \nu_n \sim N(0, \Sigma_\nu).$$

2. Distribution of the observed choice:

$$P(y_{in} = 1) = \Pr(U_{in} \geq U_{jn}, \forall j).$$

Indicators: continuous output

$$f_3(I_n|X_n, X_n^*; \alpha, \Sigma_\nu)$$

For instance:

$$I_n = m(X_n, X_n^*; \alpha) + \nu_n, \quad \nu_n \sim N(0, \sigma_{\nu_n}^2)$$

So,

$$f_3(I_n|\cdot) = \frac{1}{\sigma_{\nu_n}\sqrt{2\pi}} \exp\left(-\frac{(I_n - m(\cdot))^2}{2\sigma_{\nu_n}^2}\right)$$

Define

$$Z = \frac{I_n - m(\cdot)}{\sigma_{\nu_n}} \sim N(0,1), \quad \phi(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}$$

and

$$f_3(I_n|\cdot) = \frac{1}{\sigma_{\nu_n}}\phi(Z)$$

Indicators: discrete output

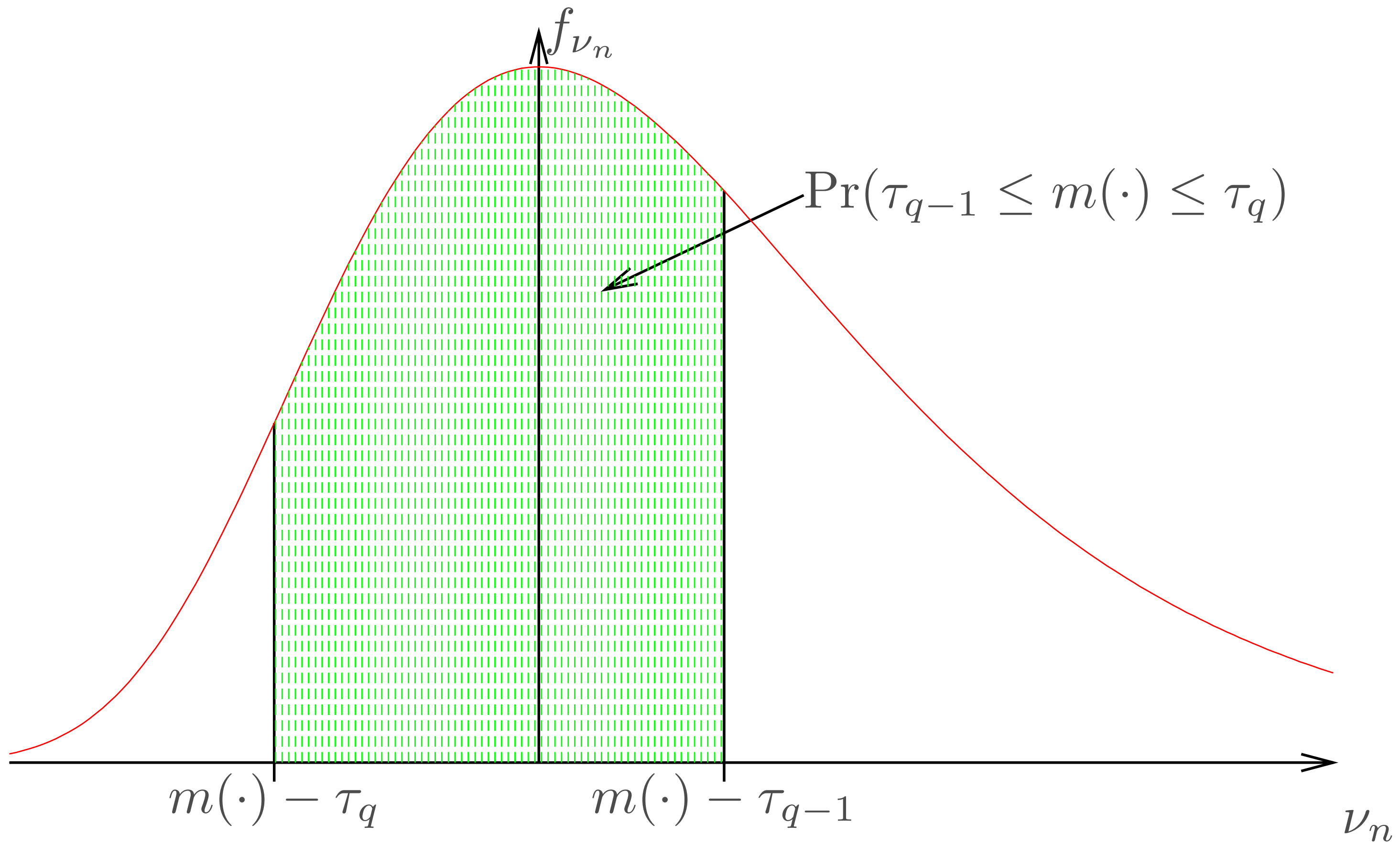
$$f_3(I_n | X_n, X_n^*; \alpha, \Sigma_\nu)$$

For instance:

$$I_n = m(X_n, X_n^*; \alpha) + \nu_n, \quad \nu_n \sim \text{Logistic(0,1)}$$

$$\begin{aligned}
P(I_n = 1) &= & \Pr(m(\cdot) \leq \tau_1) &= \frac{1}{1 + e^{-\tau_1 + m(\cdot)}} \\
P(I_n = 2) &= & \Pr(m(\cdot) \leq \tau_2) - \Pr(m(\cdot) \leq \tau_1) &= \frac{1}{1 + e^{-\tau_2 + m(\cdot)}} - \frac{1}{1 + e^{-\tau_1 + m(\cdot)}} \\
&\vdots & &\vdots \\
P(I_n = 5) &= & 1 - \Pr(m(\cdot) \leq \tau_4) &= 1 - \frac{1}{1 + e^{-\tau_4 + m(\cdot)}}
\end{aligned}$$

Indicators: discrete output



Estimation: likelihood

Assuming ω_n, ε_n and ν_n are independent, we have

$$\mathcal{L}_n(y_n, I_n | X_n; \alpha, \beta, \lambda, \Sigma_\varepsilon, \Sigma_\nu, \Sigma_\omega) =$$

$$\int_{X^*} P(y_n | X_n, X^*; \beta, \Sigma_\varepsilon) f_3(I_n | X_n, X^*; \alpha, \Sigma_\nu) f_1(X^* | X_n; \lambda, \Sigma_\omega) dX^*.$$

Maximum likelihood estimation:

$$\max_{\alpha,\beta,\lambda,\Sigma_\varepsilon,\Sigma_\nu,\Sigma_\omega} \sum_n \log\left(\mathcal{L}_n(y_n, I_n | X_n; \alpha, \beta, \lambda, \Sigma_\varepsilon, \Sigma_\nu, \Sigma_\omega)\right)$$

Source: Walker (2001)

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Software issues

- Specification of the models
- Compute integrals
- Loop on data
- Compute derivatives

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Biogeme

Current version:

- Designed for discrete choice (MEV models).
- Syntax: specific to Biogeme.
- Models are pre-implemented, with their derivatives.
- Users focus only on the utility functions.
- Additional features for random parameters, and panel data.

New version:

- Designed for any model
- Syntax: python 3.1 (case sensitive) `python.org`
- Compiler: python $\rightarrow$ C++ (likelihood and derivatives)
- Users can create the model from scratch, or use pre-implemented libraries
- Additional features for integration, and loops.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Example: a logit model

Import libraries

```
from biogeme import *
from headers import *
from logit import *
from loglikelihood import *
from statistics import *
```

Example: a logit model

Define the data file:

```
dataFile = "sample.dat"
```

Define the parameters to be estimated:

```
ASC1  =  Beta( 'ASC1' , 0 , -10000, 10000, 1)
ASC2  =  Beta( 'ASC2' , 0 , -10000, 10000, 0)
ASC3  =  Beta( 'ASC3' , 0 , -10000, 10000, 0)
ASC4  =  Beta( 'ASC4' , 0 , -10000, 10000, 0)
ASC5  =  Beta( 'ASC5' , 0 , -10000, 10000, 0)
ASC6  =  Beta( 'ASC6' , 0 , -10000, 10000, 0)
BETA1 =  Beta( 'BETA1', 0 , -10000, 10000, 0)
BETA2 =  Beta( 'BETA2', 0 , -10000, 10000, 0)
```

Example: a logit model

Define the utility functions (structural equations):

```
V1 = ASC1 + BETA1 * x11 + x12 * BETA2
V2 = ASC2 + BETA1 * x21 + BETA2 * x22
V3 = ASC3 + BETA1 * x31 + BETA2 * x32
V4 = ASC4 + BETA1 * x41 + x42 * BETA2
V5 = ASC5 + BETA1 * x51 + BETA2 * x52
V6 = ASC6 + BETA1 * x61 + BETA2 * x62
```

Assign alternatives IDs to utility functions:

```
V = {1: V1,
     2: V2,
     3: V3,
     4: V4,
     5: V5,
     6: V6}
```

Python data structure: dictionary

http://docs.python.org/py3k/tutorial/datastructures.html#dictionaries

Example: a logit model

Assign availabilities to alternatives:

```
av = {1: av1,
      2: av2,
      3: av3,
      4: av4,
      5: av5,
      6: av6}
```

Measurement equation: logit model

- use the library
- or write it yourself

Example: a logit model

Using the library:

```
prob = logit_av(V,av,Choice)
rowIterator('obsIter', Datafile(dataFile))
BIOGEME_OBJECT.ESTIMATE = Sum(log(prob),'obsIter')
```

- Iterators:
  - `Sum(formula,iterator)`,
  - `Prod(formula,iterator)`
- Function to be maximized: `BIOGEME_OBJECT.ESTIMATE`

Example: a logit model

Not using the library:

$$P(i) = \frac{e^{V_i}}{\sum_j a_j e^{V_j}} = \frac{1}{\sum_j a_j e^{V_j - V_i}}, \quad \log P(i) = -\log(\sum_j a_j e^{V_j - V_i})$$

```
chosen = Elem(V,Choice)
den = 0
for i,v in V.items() :
    den += av[i] * exp(v-chosen)
logprob = -log(den)
rowIterator('obsIter', Datafile(dataFile))
BIOGEME_OBJECT.ESTIMATE = Sum(logprob,'obsIter')
```

- `Elem(dict,index)` returns the element of a dictionary when the index varies across the sample.
- Syntax for loops from python.

Example: a latent variable

Attitude towards car: structural equation

```
omega = RandomVariable('omega')
attCar = b_cteAttCar + b_female * female + b_sigma * omega
```

Measurement equations

```
z04 = (I04 - b_alpha04 - b_lambda04 * attCar) / b_sigma04
f04 = normalpdf(z04)/ b_sigma04

z05 = (I05 - b_alpha05 - b_lambda05 * attCar) / b_sigma05
f05 = normalpdf(z05)/ b_sigma05

z10 = (I10 - b_alpha10 - b_lambda10 * attCar) / b_sigma10
f10 = normalpdf(z10)/ b_sigma10

z12 = (I12 - b_alpha12 - b_lambda12 * attCar) / b_sigma12
f12 = normalpdf(z12)/ b_sigma12
```

Example: a latent variable

Likelihood

```
condLike = P * f04 * f05 * f10 * f12
like = Integrate(condLike * normalpdf(omega),'omega')
rowIterator('obsIter', Datafile(dataFile))
BIOGEME_OBJECT.ESTIMATE = Sum(log(like),'obsIter')
```

- `Integrate` uses Gauss-Hermite integration
- If integrals of more than one dimension, simulation can be used.
- We illustrate simulation on the same example.

Example: a latent variable, with simulation

```
omega = bioNormal('omega')
condLike = P * f04 * f05 * f10 * f12
drawIterator('drawIter')
like = Sum(condLike,'drawIter')
rowIterator('obsIter', Datafile(dataFile))
BIOGEME_OBJECT.ESTIMATE = Sum(log(like),'obsIter')
```

- Sum is used here to approximate the integral
- Same syntax, with another iterator
- Other iterators exist, in particular for panel data.

Sample enumeration

- Copy the value of the estimated parameters from the file `mymodel_param.py`
- Define the quantities that must be computed

```
P1 = Integrate(condP1*density,'omega')
costElasticity = Derive(P1,'cost_1')  * cost_1 / P1
```

- Gather them in a dictionary, with appropriate labels:

```
S = {'Prob. alt. 1': P1,
     'Cost elasticity alt. 1': costElasticity}
```

- Instruct BIOGEME to simulate:

```
BIOGEME_OBJECT.SIMULATE = Enumerate(S,'obsIter')
```

Technicalities

- Alpha version is running, not distributed yet.
- Requires a GNU environment (Linux or Mac OS X)
- Exploits multiple processors
- Output in HTML, with sortable tables