New features in Biogeme

Michel Bierlaire

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

June 2, 2023

Biogeme

Recent release

- Version 3.2.11.
- Released April 19, 2023.
- `pip install biogeme`

Important features

- Sampling of alternatives.
- Segmentation.
- Catalogs.

Sampling of alternatives

Context

- Maximum likelihood estimation.
- Very large choice set $\mathcal{C}_n$ of size J_n.
- Difficult or impossible to handle.
- Idea: sample $\tilde{J}_n \leq J_n$ alternatives for each individual n.
- Sample: $\mathcal{D}_n \subseteq \mathcal{C}_n$.

Outline

Estimation of the parameters

Log likelihood maximization

$$\max_{\theta} \sum_{n=1}^{N} \ln \Pr(i_n|\mathcal{C}_n;\theta).$$

Consistent and efficient estimator.

Conditional log likelihood maximization

$$\max_{\theta} \sum_{n=1}^{N} \ln \Pr(i_n|\mathcal{D}_n,\mathcal{C}_n;\theta).$$

Consistent estimator.

Conditional log likelihood

Derivation

$$\Pr(i, \mathcal{D}|\mathcal{C};\theta) = \Pr(\mathcal{D}|i)\Pr(i|\mathcal{C};\theta).$$

Bayes' theorem:

$$\Pr(i|\mathcal{D},\mathcal{C};\theta) = \frac{\Pr(\mathcal{D}|i)\Pr(i|\mathcal{C};\theta)}{\sum_{j\in\mathcal{D}}\Pr(\mathcal{D}|j)\Pr(j|\mathcal{C};\theta)}.$$

Logit

$$\Pr(i|\mathcal{C};\theta) = \frac{e^{\mu V_i}}{\sum_{j\in\mathcal{C}} e^{\mu V_j}} = \frac{e^{\mu V_i}}{\gamma}.$$

Conditional log likelihood

Logit

$$
\begin{aligned}
\Pr(i|\mathcal{D},\mathcal{C};\theta) &= \frac{\Pr(\mathcal{D}|i)\Pr(i|\mathcal{C};\theta)}{\sum_{j\in\mathcal{D}}\Pr(\mathcal{D}|j)\Pr(j|\mathcal{C};\theta)} \\
&= \frac{\Pr(\mathcal{D}|i)\ \exp(\mu V_i)\cancel{\gamma}}{\cancel{\gamma}\sum_{j\in\mathcal{D}}\Pr(\mathcal{D}|j)\ \exp(\mu V_j)} \\
&= \frac{\exp(\mu V_i + \ln(\Pr(\mathcal{D}|i)))}{\sum_{j\in\mathcal{D}}\ \exp(\mu V_j + \ln(\Pr(\mathcal{D}|j)))}.
\end{aligned}
$$

Simple specification

Logit on $\mathcal{D}$ + correction.

Conditional log likelihood

MEV models

$$\Pr(i|\mathcal{C};\theta) = \frac{\exp(\mu V_i + \ln G_i(V_1,\ldots,V_J))}{\sum_{j\in\mathcal{C}} \exp(\mu V_j) + \ln G_j(V_1,\ldots,V_J)}$$

Exact same derivation

$$\frac{\exp(\mu V_i + \ln G_i(V_1,\ldots,V_J) + \ln(\Pr(\mathcal{D}|i)))}{\sum_{j\in\mathcal{D}} \exp(\mu V_j + \ln G_j(V_1,\ldots,V_J) + \ln(\Pr(\mathcal{D}|j))}.$$

But...

$$G_i(V_1,\ldots,V_J)$$

involves all alternatives. Must be approximated.

Nested logit model

MEV term

$$\ln G_i = (\frac{\mu}{\mu_m} - 1)(\ln \sum_{j \in \mathcal{C}_m} \exp(\mu_m V_j)) + \ln \mu + (\mu_m - 1) V_i.$$

Approximation

[Guevara and Ben-Akiva, 2013]

$$\sum_{j \in \mathcal{C}_m} \exp(\mu_m V_j) \approx \sum_{j \in \mathcal{C}_m \cap D} w_j \exp(\mu_m V_j)$$

where

$$w_j = \frac{1}{\Pr(\mathcal{D}|j)}.$$

Outline

Sampling protocol

Procedure

- Partition the choice set into K segments of size R_k: $J = \sum_k R_k$.
- Decide the number r_k of alternatives to select in each segment.
- Size of D: $\sum_k r_k$.
- Alternative i is in segment $k(i)$.
- Randomly draw $r_{k(i)} - 1$ alternatives among the non chosen ones in segment $k(i)$, and add i: $\mathcal{D}_{k(i)}$.
- Randomly draw r_k alternatives in each segment k, $k \neq k(i)$: $\mathcal{D}_k$.
- The sample is composed of the chosen alternative and all draws: $\mathcal{D} = \cup_k \mathcal{D}_k$.

Sampling protocol

Correction term

$$\Pr(\mathcal{D}|i) = \binom{R_{k(i)}-1}{r_{k(i)}-1} \prod_{k=2}^{K} \binom{R_k}{r_k}$$

$$= \frac{R_{k(i)}}{r_{k(i)}} \prod_{k=1}^{K} \binom{R_k}{r_k}$$

$$= \frac{1}{\Pr(i)} \Pr(\mathcal{D})$$

Sampling protocol

Logit

$$\begin{aligned}
\Pr(i|\mathcal{D},\mathcal{C};\theta) &= \frac{\exp(\mu V_i + \ln(\Pr(\mathcal{D}|i)))}{\sum_{j\in\mathcal{D}} \exp(\mu V_j + \ln(\Pr(\mathcal{D}|j)))} \\
&= \frac{\exp(\mu V_i + \cancel{\ln(\Pr(\mathcal{D}))} + \ln R_{k(i)} - \ln r_{k(i)})}{\sum_{j\in\mathcal{D}} \exp(\mu V_j + \cancel{\ln(\Pr(\mathcal{D}))} + \ln R_{k(j)} - \ln r_{k(j)})} \\
&= \frac{\exp(\mu V_i + \ln R_{k(i)} - \ln r_{k(i)})}{\sum_{j\in\mathcal{D}} \exp(\mu V_j + \ln R_{k(j)} - \ln r_{k(j)})}
\end{aligned}$$

Sampling protocol

MEV terms

$$
\begin{aligned}
\sum_{j \in \mathcal{C}_m} \exp(\mu_m V_j) &\approx \sum_{j \in \mathcal{C}_m \cap D} w_j \exp(\mu_m V_j) \\
&= \frac{1}{\Pr(\mathcal{D})} \sum_{j \in \mathcal{C}_m \cap D} \Pr(i) \exp(\mu_m V_j) \\
&= \frac{1}{\Pr(\mathcal{D})} \sum_{j \in \mathcal{C}_m \cap D} \frac{r_{k(i)}}{R_{k(i)}} \exp(\mu_m V_j) \\
&= \frac{1}{\Pr(\mathcal{D})} \sum_{j \in \mathcal{C}_m \cap D} \exp(\mu_m V_j - \ln R_{k(j)} + \ln r_{k(j)}),
\end{aligned}
$$

where

$$
w_j = \frac{1}{\Pr(\mathcal{D}|j)}.
$$

In practice

Correction term

For each i, we need

$$\ln R_{k(i)} - \ln r_{k(i)}.$$

Bibliography I

Guevara, C. A. and Ben-Akiva, M. (2013).
Sampling of alternatives in multivariate extreme value (MEV) models.
Transportation Research Part B, 48:31– 52.