Empirical investigation of sampling of alternatives of Multivariate Extreme Value (MEV) migration aspiration discrete choice models

Evangelos Paschalidis[1], Andreas B. Vortisch[2], Michel Beine[2], Michel Bierlaire[1]

[1]Transport and Mobility Laboratory (TRANSP-OR)
School of Architecture, Civil and Environmental Engineering
École Polytechnique Fédérale de Lausanne

[2]Department of Economics and Management University of Luxembourg

16th workshop on Discrete Choice Models, June 6 - 8, 2024, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Context - Sampling of alternatives

- The problem:
  - Model estimation is challenging with large number of alternatives
  - The problem escalates for large number of observations
  - Specifications such as MEV or mixtures can pose extra challenge
- The solution: Sampling of alternatives (work on a subset of alternatives)

Context - Sampling of alternatives

- McFadden (1978) proposed a consistent estimation under sampling of alternatives with a simple correction to the log-likelihood function
- Guevara & Ben-Akiva (2013) extended the sampling of alternatives technique in MEV models
- Guevara & Ben-Akiva (2013) extended the sampling of alternatives technique in logit mixtures models

Migration and Discrete Choice Models (MIGDCM) project

Objective: Use of advanced discrete choice models to capture migration aspirations

- Model specification
- Model transferability
- Relationship between aspiration and preparation plans
- Sampling of alternatives

Why sampling of alternatives?

- Global choice set: 200+ countries
- Long list of attributes per destination
- Large number of observations
- Problems increase for more complex model specifications (e.g. CNL)
- Our goal: Investigate the use of sampling of alternatives in the context of migration aspiration discrete choice models

Theoretical background: Sampling of alternatives

- We assume C the full choice set, containing J alternatives (same for all individuals)
- Choice model $Pr(i|C;\theta)$

$$\max_{\theta} \sum_{n=1}^{N} \ln Pr(i|C;\theta)$$

- We can instead model on a subset D_n of C (conditional maximum likelihood estimator)

$$\max_{\theta} \sum_{n=1}^{N} \ln Pr(i|D_n, C;\theta)$$

- Consistent but not asymptotically efficient

Theoretical background

- The sum must not depend on C
- Using Bayes' Theorem and Total Probability theorem

$$Pr(i_n|D_n, C; \theta) = Pr(D_n|i_n)Pr(i_n|C; \theta)$$

$$Pr(i_n|D_n, C; \theta) = \frac{Pr(D_n|i_n)Pr(i_n|C; \theta)}{\sum_{j \in D_n} Pr(D_n|j_n)Pr(j_n|C; \theta)}$$

Theoretical background - Logit model

- Let us consider:

$$Pr(i_n|C;\theta) = \frac{e^{\mu V_{i_n}}}{\sum_{j\in C} e^{\mu V_{j_n}}} = \frac{e^{\mu V_{i_n}}}{\gamma}$$

- Then we have:

$$Pr(i_n|D_n, C;\theta) = \frac{Pr(D_n|i_n)Pr(i_n|C;\theta)}{\sum_{j\in D_n} Pr(D_n|j_n)Pr(j_n|C;\theta)}$$

$$= \frac{Pr(D_n|i_n)\exp(\mu V_{i_n})\cancel{\gamma}}{\cancel{\gamma}\sum_{j\in D_n} Pr(D_n|j_n)\exp(\mu V_{j_n})}$$

$$= \frac{\exp(\mu V_{i_n} + \ln Pr(D_n|i_n))}{\sum_{j\in D_n} \exp(\mu V_{j_n} + \ln Pr(D_n|j_n))}$$

Theoretical background - MEV models

- MEV model probability with a generating function G:

$$Pr(i_n|C;\theta) = \frac{\exp(\mu V_{i_n} + \ln G_{i_n}(V_1, ..., V_j))}{\sum_{j \in C} \exp(\mu V_{j_n} + \ln G_{j_n}(V_1, ..., V_j))}$$

- We can expand to MEV models as

$$Pr(i_n|D_n, C;\theta) = \frac{\exp(\mu V_{i_n} + \ln G_{i_n}(V_1, ..., V_j) + \ln Pr(D_n|i_n))}{\sum_{j \in D_n} \exp(\mu V_{j_n} + \ln G_{j_n}(V_1, ..., V_j) + \ln Pr(D_n|j_n))}$$

- But $G_{i_n}(V_1, ..., V_j)$ involves all alternatives - needs to be approximated

Sampling of alternatives: Nested logit model

- NL generating function:

$$\ln G_i = \left(\frac{\mu}{\mu_m} - 1\right)\left(\ln \sum_{j \in C_m} \exp(\mu_m V_j)\right) + \ln \mu + (\mu_m - 1) V_i$$

- From Guevarra & Ben-Akiva (2013):

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

- where:

$$w_j = \frac{1}{Pr(j)}$$

Sampling of alternatives: Cross-nested logit model

- CNL generating function:

$$G_i = \mu \sum_{m=1}^{M} \alpha_{im}^{\frac{\mu_m}{\mu}} \exp((\mu_m - 1) V_{in}) \left(\sum_{j \in C} \alpha_{jm}^{\frac{\mu_m}{\mu}} \exp(\mu_m V_{jn}) \right)^{\frac{\mu}{\mu_m} - 1}$$

- From Guevarra & Ben-Akiva (2013):

$$\sum_{j \in C_m} \alpha_{jm}^{\frac{\mu_m}{\mu}} \exp(\mu_m V_j) \approx \sum_{j \in C_m \cap D} w_j \alpha_{jm}^{\frac{\mu_m}{\mu}} \exp(\mu_m V_j)$$

- where:

$$w_j = \frac{1}{Pr(j)}$$

Sampling procedure in Biogeme - Utility correction

1. Partition the full choice set into K segments (strata) of size R_k : $J = \sum_{k=1}^{K} R_k$
2. Define a number r_k which represents the number of alternatives to be sampled from each stratum, $D_n : \sum_{k=1}^{K} r_k$
3. Denote $k(i)$ the stratum containing the chosen alternative i
4. Randomly draw $r_{k_i} - 1$ alternatives among the non chosen ones in stratum $k(i)$ and add i to obtain $D_k(i)$
5. Randomly draw r_k alternatives in each stratum k, $k \neq k(i)$ to obtain D_k
6. Compute $\ln Pr(D|i) = \ln R_{k(i)} - \ln r_{k(i)}$ ($\frac{R_k(i)}{r_k(i)} \propto \Pr(D|i)$)

Sampling procedure in Biogeme - G function (re-sampling possible)

- Same process as for the utility correction.
- The chosen alternative does not play any role in the sampling procedure.
- It is not necessary to partition the full choice set; alternatives that are alone in a nest do not contribute in the generating function and can be excluded.
- Every nest must be represented in the sampling procedure.
- We calculate: $w_{jn} = \frac{1}{\Pr(j)} = \frac{R_k(j)}{r_k(j)}$.

Case study - Migration aspiration in Mexico

Beine et al. (Working paper): *The Impact of a Possible Trump Reelection on Mexican Immigration Pressures in Alternative Countries*

- Migration aspirations in Mexico
- Implementation of logit, NL, and CNL models for migration aspiration (origin: Mexico)
- Sensitivity to moving to USA under different scenarios (i.e. Donald J. Trump sensitivity...)
- Choice set size 96 alternatives: (69 chosen + 27 non-chosen)
- 10,081 observations were considered for model estimation.
- 1,821 individuals aspired to move (approx. 18%)

Data

- Gallup World Poll (GWP) surveys on migration aspirations from Mexico over the period 2007-2019
  - Migration aspirations (yes/no) and destination choice
  - Socio-demographic characteristics (gender, age, skill level, income, number of children, area of residence)

- Destination-specific variables
  - GDP
  - Mexican diaspora
  - Distance from Mexico
  - Language (English or not)
  - Contiguity

Deviations from the original study

Original study	Sampling of alternatives
10'081 observations	1'821 observations (stayers excluded)
NL nests: 1. foreign, 2. stay	NL nests: 1.OECD/Schengen & non-English speaking, 2. English speaking
CNL nests: 1. OECD, 2. Schengen, 3. English speaking	CNL nests: 1. OECD, 2. Schengen, 3. English speaking

Sampling protocols

- Random sampling: Each alternative has an equal probability to be selected.
- Importance sampling: Sampling incorporating stratification of alternatives.

Insights on choices

Destination	Freq.	%	% Cumul.
USA	750	41.19%	41.19%
Canada	255	14.00%	55.19%
Spain	133	7.30%	62.49%
Germany	122	6.70%	69.19%
France	84	4.61%	73.81%
Italy	43	2.36%	76.17%
Brazil	41	2.25%	78.42%
Japan	32	1.76%	80.18%
China	32	1.76%	81.93%
United Kingdom	29	1.59%	83.53%
Switzerland	26	1.43%	84.95%
Australia	25	1.37%	86.33%
Cuba	23	1.26%	87.59%
Albania	21	1.15%	88.74%
Russia	18	0.99%	89.73%

Sampling protocols - Utility sampling

- Protocol 1: Random sampling
- Protocols 2 & 3: 2 strata (OECD, nonOECD)*
- Protocols 4 & 5: 4 strata (OECD, nonOECD, USA, Canada)

*In the sample there are 35 OECD and 61 non-OECD countries

Sampling protocols - MEV sampling

- Protocol 1: Random sampling
- Protocols 2 & 4: 2 strata (OECD & Schengen (no English), English)
- Protocols 3 & 5: 4 strata (OECD & Schengen (no English), English, USA, Canada)

Sample sizes - Estimation

- Utility correction term: 20, 40, 60 alternatives
- MEV correction term: 20, 40, and 56 (all) alternatives sampled for the generating function
- Implementation for NL model and CNL model
- 100 estimations per sampling protocol (BFGS)

Sample sizes - Utility sampling

	Protocol 2 & 3			Protocol 4 & 5		
	20	**40**	**60**	**20**	**40**	**60**
OECD	10	20	30	8	18	28
non-OECD	10	20	30	10	20	30
USA	-	-	-	1	1	1
CAN	-	-	-	1	1	1

Sample sizes - MEV sampling

	Protocol 2 & 4			Protocol 3 & 5		
	20	**40**	**56**	**20**	**40**	**56**
Group 1	10	20	30	10	20	-
Group 2	10	20	26	8	18	-
USA	-	-	-	1	1	-
CAN	-	-	-	1	1	-

Analysis - Metrics

- true: the "true" value β^*
- mean: the empirical mean $\widehat{\beta_k}$ over the 100 estimations
- stdev: the empirical standard deviation $\widehat{\sigma_k}$ over the 100 estimations
- ttest-true: the ratio $\frac{\beta^* - \widehat{\beta_k}}{\widehat{\sigma_k}}$
- lowBound: $\beta^* - \Phi^{-1}(1 - 0.125)\widehat{\sigma_k}$
- upBound: $\beta^* + \Phi^{-1}(1 - 0.125)\widehat{\sigma_k}$
- count: empirical coverage, number of repetitions for which the estimator was within a normal 75% confidence interval, that is the percentage of estimates lying in the interval [lowBound, upBound]

Results - Random sampling (U60-MEV56) [NL model]

betas_names	betas_original	beta_mean	beta_std	t_ratio	lowerbound	upperbound	counts_sum
MU_nest1	0.666	0.688	0.013	-1.69	0.650	0.681	27
MU_nest2	0.478	0.498	0.015	-1.28	0.460	0.496	48
beta_GDP_eduCollege	0.588	0.608	0.010	-2.03	0.577	0.599	20
beta_GDP_eduElementary	-1.765	-1.838	0.082	0.90	-1.858	-1.671	55
beta_GDP_eduSecondary	-0.602	-0.541	0.096	-0.63	-0.713	-0.491	67
beta_disapproval_eduCollege	-0.897	-0.933	0.046	0.78	-0.950	-0.844	66
beta_disapproval_eduElementary	0.117	0.121	0.007	-0.62	0.109	0.125	64
beta_disapproval_eduSecondary	0.164	0.151	0.006	2.27	0.157	0.170	11
beta_englishSpeakingCountries	0.139	0.137	0.003	0.43	0.135	0.142	75
beta_logdiaspora_eduCollege	-0.552	-0.574	0.013	1.60	-0.568	-0.537	29
beta_logdiaspora_eduElementary	0.424	0.437	0.005	-2.72	0.418	0.429	3
beta_logdiaspora_eduSecondary	-0.191	-0.158	0.031	-1.02	-0.227	-0.154	55
beta_logdist	1.605	1.560	0.021	2.14	1.580	1.629	14
beta_logpopul	0.430	0.447	0.010	-1.79	0.419	0.441	23
beta_oecdCountries	2.188	2.144	0.019	2.34	2.166	2.210	10
beta_schengenCountries	1.179	1.181	0.008	-0.18	1.170	1.188	76

Results - Random sampling (U60-MEV56) [CNL model]

betas_names	betas_original	beta_mean	beta_std	t_ratio	lowerbound	upperbound	counts_sum
beta_GDP_eduCollege	0.533	0.561	0.013	-2.11	0.518	0.548	15
beta_GDP_eduElementary	0.372	0.393	0.012	-1.75	0.359	0.386	24
beta_GDP_eduSecondary	0.488	0.510	0.009	-2.54	0.478	0.498	9
beta_disapproval_eduCollege	-1.335	-1.408	0.069	1.05	-1.414	-1.256	53
beta_disapproval_eduElementary	-0.593	-0.567	0.076	-0.34	-0.681	-0.506	70
beta_disapproval_eduSecondary	-0.712	-0.756	0.041	1.07	-0.759	-0.664	51
beta_englishSpeakingCountries	0.100	0.103	0.006	-0.49	0.093	0.106	71
beta_logdiaspora_eduCollege	0.152	0.144	0.006	1.39	0.146	0.159	43
beta_logdiaspora_eduElementary	0.120	0.120	0.003	0.13	0.117	0.123	77
beta_logdiaspora_eduSecondary	-0.419	-0.443	0.013	1.86	-0.434	-0.405	29
beta_logdist	0.324	0.339	0.004	-3.58	0.319	0.329	3
beta_logpopul	0.308	0.318	0.022	-0.45	0.283	0.333	69
beta_oecdCountries	1.855	1.819	0.016	2.23	1.836	1.874	15
beta_schengenCountries	0.713	0.716	0.009	-0.33	0.702	0.723	74
param_MU_English	1.290	1.288	0.009	0.24	1.279	1.300	69
param_MU_OECD	2.135	2.100	0.015	2.38	2.118	2.152	13
param_MU_Schengen	3.105	2.999	0.036	2.91	3.063	3.147	3

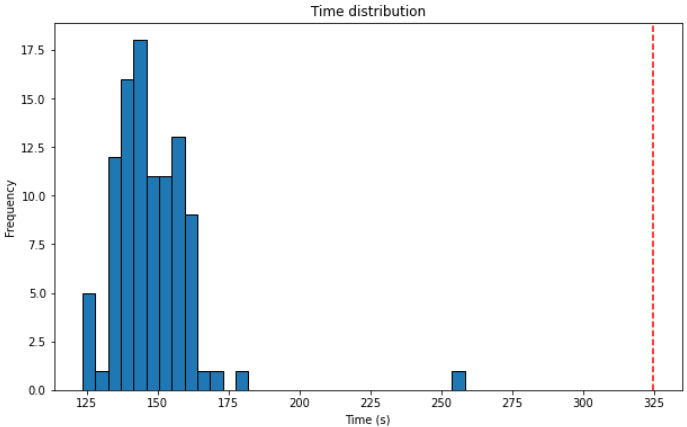
Results - t-ratio [NL model (1/2)]

	P2 U20 MEV56	P2 U40 MEV56	P2 U60 MEV56	P2 U60 MEV40	P3 U60 MEV40
MU_nest1	-2.18	-1.27	-0.60	-2.43	-2.09
MU_nest2	-1.65	-1.36	-0.67	-1.22	-0.86
beta_GDP_eduCollege	-2.51	-1.73	-0.81	-2.15	-1.85
beta_GDP_eduElementary	1.23	0.77	0.58	1.34	1.13
beta_GDP_eduSecondary	-0.82	-0.81	-0.12	-0.55	-0.25
beta_disapproval_eduCollege	0.82	0.61	0.22	0.36	0.25
beta_disapproval_eduElementary	-1.20	-0.90	-0.51	-1.71	-1.39
beta_disapproval_eduSecondary	3.38	2.53	0.99	-0.70	-0.49
beta_englishSpeakingCountries	0.57	0.45	0.18	-2.13	-1.42
beta_logdiaspora_eduCollege	2.17	1.58	0.80	1.67	1.41
beta_logdiaspora_eduElementary	-4.31	-1.96	-0.87	-2.80	-2.30
beta_logdiaspora_eduSecondary	-1.42	-0.68	-0.43	1.76	1.18
beta_logdist	3.14	1.57	0.78	3.79	2.95
beta_logpopul	-2.24	-1.52	-0.66	0.38	-0.01
beta_oecdCountries	3.82	1.74	0.90	3.63	2.83
beta_schengenCountries	-0.18	-0.01	-0.17	2.59	2.09

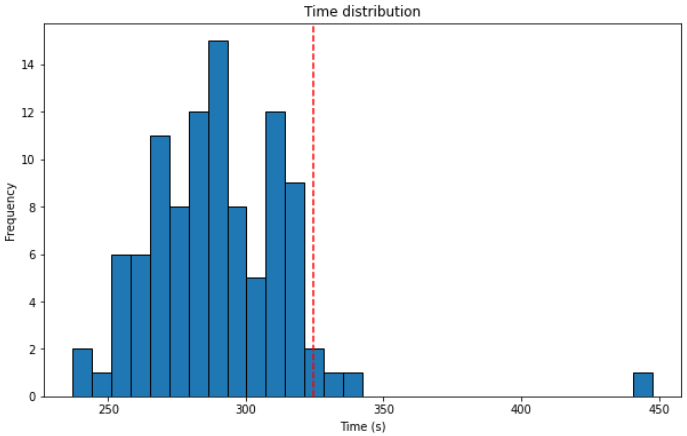
Results - t-ratio [NL model (2/2)]

	P4 U20 MEV56	P4 U20 MEV40	P4 U40 MEV56	P4 U40 MEV40	P4 U60 MEV56	P4 U60 MEV40	P5 U40 MEV40	P5 U60 MEV40
MU_nest1	-0.56	-1.80	-0.15	-2.09	0.01	-2.38	-0.85	-0.52
MU_nest2	-0.12	-0.56	-0.06	-0.54	0.03	-0.86	0.09	0.26
beta_GDP_eduCollege	-0.57	-1.83	-0.38	-1.80	0.08	-2.05	-0.61	-0.33
beta_GDP_eduElementary	0.19	1.16	0.02	1.86	-0.02	2.00	0.89	0.99
beta_GDP_eduSecondary	0.20	-0.07	-0.02	-0.22	-0.17	-0.54	0.22	0.22
beta_disapproval_eduCollege	0.03	0.37	-0.17	0.55	-0.11	0.60	0.66	1.07
beta_disapproval_eduElementary	1.65	-0.02	0.61	-1.79	0.25	-2.37	0.32	-0.28
beta_disapproval_eduSecondary	1.36	-0.15	0.59	-1.82	0.23	-2.18	0.37	-0.11
beta_englishSpeakingCountries	1.90	0.01	0.87	-1.87	0.18	-2.49	0.64	-0.14
beta_logdiaspora_eduCollege	0.82	1.51	0.49	1.52	0.17	1.45	0.70	0.15
beta_logdiaspora_eduElementary	-0.16	-2.36	0.19	-2.60	0.34	-2.60	-0.99	-1.15
beta_logdiaspora_eduSecondary	-1.05	0.95	-0.55	1.96	-0.09	2.22	-2.51	-2.73
beta_logdist	-0.18	2.23	-0.24	3.31	-0.26	3.81	2.35	3.13
beta_logpopul	-0.23	0.34	0.02	0.54	0.12	0.89	-0.61	-0.78
beta_oecdCountries	-0.62	2.22	-0.41	3.26	-0.43	3.56	1.91	2.11
beta_schengenCountries	-0.56	2.30	-0.41	2.93	-0.10	2.88	0.20	0.45

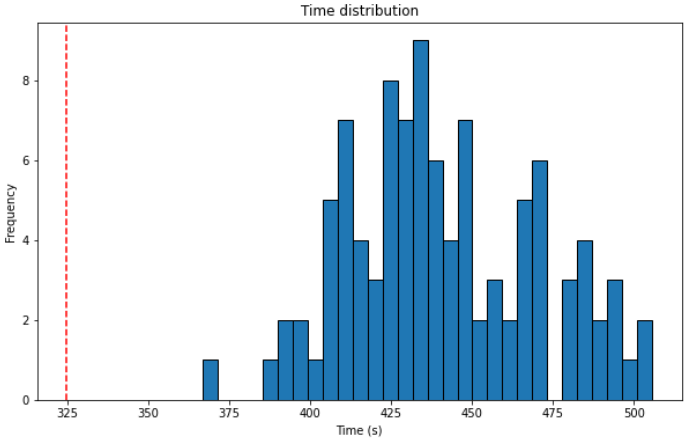
Estimation time - P4 U20 MEV56



Estimation time - P4 U40 MEV56



Estimation time - P4 U60 MEV56



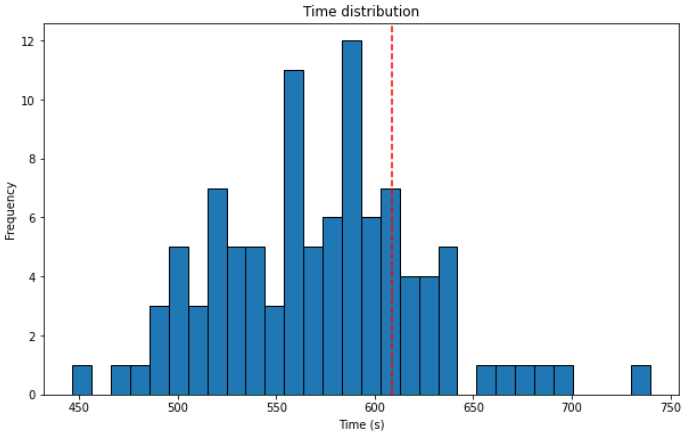
Results - t-ratio [CNL model (1/2)]

	P1 U60 MEV56	P2 U20 MEV56	P2 U40 MEV56	P2 U60 MEV56	P2 U60 MEV40	P3 U60 MEV40
beta_GDP_eduCollege	-2.11	-3.67	-1.99	-0.81	-2.50	-2.44
beta_GDP_eduElementary	-1.75	-2.52	-1.30	-0.69	-0.90	-0.87
beta_GDP_eduSecondary	-2.54	-3.82	-2.33	-1.14	-2.29	-1.98
beta_disapproval_eduCollege	1.05	1.44	1.22	0.70	1.70	1.75
beta_disapproval_eduElementary	-0.34	-0.42	-0.24	-0.01	-0.33	-0.22
beta_disapproval_eduSecondary	1.07	2.08	1.29	0.54	1.08	0.89
beta_englishSpeakingCountries	-0.49	-0.89	-0.66	-0.38	-1.41	-1.16
beta_logdiaspora_eduCollege	1.39	2.78	1.55	0.70	0.04	0.36
beta_logdiaspora_eduElementary	0.13	0.09	0.19	0.19	-1.21	-0.82
beta_logdiaspora_eduSecondary	1.86	3.23	2.05	1.08	1.48	1.55
beta_logdist	-3.58	-5.36	-2.76	-1.19	-3.39	-3.06
beta_logpopul	-0.45	-0.39	-0.31	-0.19	-0.21	-0.73
beta_oecdCountries	2.23	3.18	1.79	0.68	4.77	4.59
beta_schengenCountries	-0.33	-0.15	-0.35	-0.22	-0.76	-0.82
param_MU_English	0.24	0.66	0.22	-0.05	1.78	1.24
param_MU_OECD	2.38	3.22	1.70	0.73	4.39	3.89
param_MU_Schengen	2.91	5.42	2.56	1.11	2.74	2.30

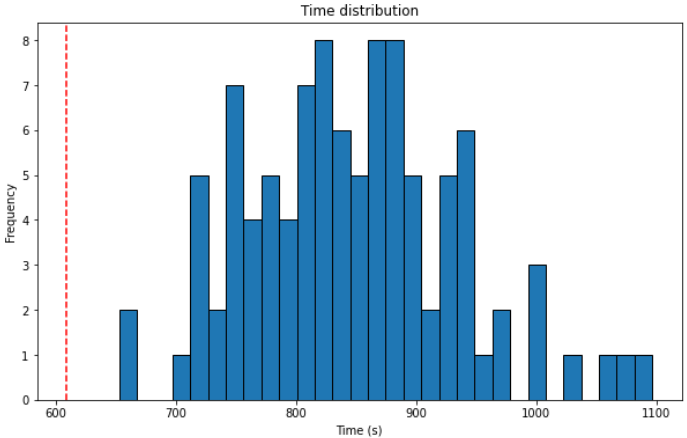
Results - t-ratio [CNL model (2/2)]

	P4 U20 MEV56	P4 U40 MEV56	P4 U60 MEV56	P4 U60 MEV40	P5 U60 MEV40	P5 U40 MEV40
beta_GDP_eduCollege	-0.87	-0.43	-0.05	-2.43	-1.27	-1.17
beta_GDP_eduElementary	0.22	-0.03	0.05	-0.64	-0.14	0.09
beta_GDP_eduSecondary	-0.70	-0.39	-0.09	-1.90	-0.63	-0.66
beta_disapproval_eduCollege	0.19	0.13	0.00	2.50	1.90	1.26
beta_disapproval_eduElementary	0.39	0.18	-0.06	-0.77	0.32	0.28
beta_disapproval_eduSecondary	0.38	0.35	-0.01	1.39	2.13	1.78
beta_englishSpeakingCountries	1.76	0.96	0.20	-1.76	0.33	0.90
beta_logdiaspora_eduCollege	1.35	0.85	0.20	-0.69	0.22	0.54
beta_logdiaspora_eduElementary	2.39	1.10	0.28	-1.69	0.08	0.73
beta_logdiaspora_eduSecondary	0.90	0.54	0.27	1.18	1.56	1.46
beta_logdist	-0.75	-0.21	0.19	-2.91	-1.87	-1.72
beta_logpopul	-1.56	-0.78	-0.06	-0.21	-2.24	-1.98
beta_oecdCountries	0.17	0.06	-0.30	4.98	2.74	2.21
beta_schengenCountries	-0.31	-0.12	-0.04	-0.64	1.53	0.86
param_MU_English	-0.73	-0.51	-0.16	1.77	0.69	0.35
param_MU_OECD	0.35	0.08	-0.35	4.38	2.76	2.54
param_MU_Schengen	0.18	-0.12	-0.33	2.19	1.88	2.00

Estimation time - P4 U40 MEV56



Estimation time - P4 U60 MEV56



Estimation times - Full choice set

	"Standard" code	Sampling code
NL	00:06:02	00:10:34
CNL	00:10:09	00:19:31

Findings

- Sampling works well only when all alternatives are sampled in the generating function.
- Sampling works better for some sampling protocols.
- Time gains for NL and CNL models only for lower sample sizes.

Next steps

- Try data sets with more alternatives.
- Try more complex model specifications (more nests?).
- Investigate alternative sampling protocols.

Questions??