Improving the efficiency of individualized designs for the mixed logit choice model by including covariates

Marjolein CRABBE
Martina VANDEBROEK

DEMA 2011
August 30 - September 2

Outline

- Introduction
- Analysis of the mixed logit choice model with covariates
- Including covariates in experimental design for the mixed logit choice model
- Simulation study
- Conclusions

Introduction

- In choice-based conjoint studies, not only the attributes of the product profiles in the choice sets, but also covariates may influence respondents' choice behavior
  - Demographics (age, gender, ...)
  - Socio-economic data (income level, employment, ...)
  - Other individual-specific characteristics (brand or store loyalty, ...)
- Taking choice related respondent characteristics into account in the setup and analysis of the discrete choice experiment to increase the accuracy of the parameter estimates

Introduction

- The mixed logit or random-effects discrete choice model to analyze choice data
- Previous work: the inclusion of covariates in the random-effects distribution to estimate the mixed logit choice model
- This research incorporates covariates in the construction of efficient individualized designs for the mixed logit choice model

$\Rightarrow$ Can we improve the accuracy of the estimates for the individual-specific partworths in the mixed logit choice model by taking covariates into account in both design and estimation?

Analysis of the mixed logit choice model with covariates

- Hierarchical model with two levels
  - Lower respondent level
    - → Models individual choice behavior by the conditional logit (CL) model
  - Upper population level
    - → Models preference heterogeneity in the population by assuming a random-effects distribution over the individual partworths

Analysis of the mixed logit choice model with covariates

Lower respondent level

- Each person is assigned an individual-specific parameter vector $\boldsymbol{\beta}_n$, constant over all choice sets
- Conditional on $\boldsymbol{\beta}_n$, the probability that individual n chooses alternative k in choice set s (CL model)

$$p_{ksn}(\boldsymbol{\beta}_n) = \frac{\exp(\mathbf{x}'_{ksn}\boldsymbol{\beta}_n)}{\sum_{i=1}^{K}\exp(\mathbf{x}'_{isn}\boldsymbol{\beta}_n)}$$

- The likelihood of respondent n's series of choices $\mathbf{y}_n^S$ for the S choice sets in the experimental design

$$L(\boldsymbol{\beta}_n|\mathbf{y}_n^S, \mathbf{X}_n^S) = \prod_{s=1}^{S}\prod_{k=1}^{K}(p_{ksn}(\boldsymbol{\beta}_n))^{y_{ksn}}$$

Analysis of the mixed logit choice model with covariates

Upper population level

- We assume the individual partworths depend on covariates

$$\boldsymbol{\beta}_n = \boldsymbol{\Theta}\mathbf{z}_n + \boldsymbol{\xi}_n$$

  - $q \times 1$ vector $\mathbf{z}_n$ with covariates for respondent n
  - $p \times q$ matrix $\boldsymbol{\Theta}$ with regression parameters
  - $\mathcal{N}(\boldsymbol{\xi}_n|\mathbf{0}, \boldsymbol{\Sigma})$ a p-variate normal distribution

- The individual-specific partworths follow a multivariate normal distribution $\mathcal{N}(\boldsymbol{\beta}_n|\boldsymbol{\Theta}\mathbf{z}_n, \boldsymbol{\Sigma})$
- The unconditional likelihood of respondent n's $\mathbf{y}_n^S$

$$L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}|\mathbf{y}_n^S, \mathbf{X}_n^S, \mathbf{z}_n) = \int L(\boldsymbol{\beta}_n|\mathbf{y}_n^S, \mathbf{X}_n^S)\ \phi(\boldsymbol{\beta}_n|\boldsymbol{\Theta}\mathbf{z}_n, \boldsymbol{\Sigma})\ d\boldsymbol{\beta}_n$$

Analysis of the mixed logit choice model with covariates

Mixed logit choice model without covariates

$$U_{ksn} = \mathbf{x}'_{ksn}\boldsymbol{\beta}_n + \varepsilon_{ksn}$$
$$\boldsymbol{\beta}_n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Mixed logit choice model with covariates

$$U_{ksn} = \mathbf{x}'_{ksn}\boldsymbol{\beta}_n + \varepsilon_{ksn}$$
$$\boldsymbol{\beta}_n \sim \mathcal{N}(\boldsymbol{\Theta}\mathbf{z}_n, \boldsymbol{\Sigma})$$

Including covariates in experimental design for the mixed logit choice model

- Individually adapted sequential Bayesian conjoint choice designs (Yu et al. 2011)
- Superior to aggregate designs due to preference heterogeneity in the population
- Based on two-level structure of the mixed logit choice model
  - $\longrightarrow$ Individual choice behavior modeled by the conditional logit model
- Two stages
  - Initial static stage
  - Adaptive sequential stage

Including covariates in experimental design for the mixed logit choice model

Initial static stage

- Construction of an individual initial Bayesian $\mathcal{D}$-efficient design $\mathbf{X}_n^{S_1}$ with S_1 choice sets for each respondent
- Minimizing the expectation of the $\mathcal{D}$-error over a prior distribution of the model parameters
- Multivariate normal prior $\mathcal{N}(\boldsymbol{\beta}_n|\boldsymbol{\Theta}_0\mathbf{z}_n, \boldsymbol{\Sigma}_0)$
  - Covariate values for individual n in vector $\mathbf{z}_n$
  - Prior values for hyperparameters $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Sigma}_0$ obtained from a pilot study, previous experiments or expert knowledge

Including covariates in experimental design for the mixed logit choice model

Adaptive sequential stage

- Initial experiment $\mathbf{X}_n^{S_1}$ for individual n and corresponding choices $\mathbf{y}_n^{S_1}$
- Bayesian update of prior information $q(\boldsymbol{\beta}_n|\mathbf{y}_n^{S_1}, \mathbf{X}_n^{S_1}, \mathbf{z}_n, \boldsymbol{\Theta}_0, \boldsymbol{\Sigma}_0)$

$$= \frac{L(\boldsymbol{\beta}_n|\mathbf{y}_n^{S_1}, \mathbf{X}_n^{S_1})\ \phi(\boldsymbol{\beta}_n|\boldsymbol{\Theta}_0\mathbf{z}_n, \boldsymbol{\Sigma}_0)}{\int L(\boldsymbol{\beta}_n|\mathbf{y}_n^{S_1}, \mathbf{X}_n^{S_1})\ \phi(\boldsymbol{\beta}_n|\boldsymbol{\Theta}_0\mathbf{z}_n, \boldsymbol{\Sigma}_0)\ d\boldsymbol{\beta}_n}$$

Including covariates in experimental design for the mixed logit choice model

Adaptive sequential stage

- Consider all possible candidate sets for $\mathbf{x}_n^{S_1+1}$
- The additional choice set is obtained by minimizing the expected $\mathcal{D}$-error of the combined design $(\mathbf{X}_n^{S_1}, \mathbf{x}_n^{S_1+1})$ over the updated prior distribution $q(\boldsymbol{\beta}_n | \mathbf{y}_n^{S_1}, \mathbf{X}_n^{S_1}, \mathbf{z}_n, \boldsymbol{\Theta}_0, \boldsymbol{\Sigma}_0)$
- Recurring process of updating an individual's prior information by means of its observed choices and sequentially adding efficient choice sets

Simulation study

Aim

- Comparing the performance of different design and estimation strategies in obtaining accurate individual-level parameter estimates and predictions and verifying whether the incorporation of covariates in design and/or estimation is valuable
- Four different design and estimation combinations

		Design	*Estimation*
1	*C-C*	IASB with covariates	covariates
2	*NC-C*	IASB without covariates	covariates
3	*NC-NC(I)*	IASB without covariates	no covariates
4	*NC-NC(O)*	single nearly orthogonal	no covariates

- Designs of type $3^3/3/16$
- For the IASB designs, five (S_1) choice sets in initial designs

Simulation study

Setup

- Pilot study
    1. The 250 respondents from the main study are also used in the pilot study
    2. 100 additional respondents are used in the pilot study (different from the 250 in the main study)
- True choice behavior
    - A. Influenced by the covariate(s)
    - B. Not influenced by the covariate(s)

	Pilot study	*Choice behavior*
I	1. 250 main resp	A. influenced by covariate(s)
II		B. not influenced by covariate(s)
III	2. 100 additional resp	A. influenced by covariate(s)
IV		B. not influenced by covariate(s)

- Discussion of the results for one binary covariate
- Similar results for the two-covariate case

Simulation study

Performance measures

- Estimation accuracy
  - The root mean squared estimation error

$$RMSE_{\beta} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*)'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*)}$$

  - The percentage of respondents for which the approach provides the smallest individual estimation error
- Prediction accuracy
  - Design including all possible choice sets with three alternatives (2925×3 profiles)
  - The root mean squared prediction error

$$RMSE_{p} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\mathbf{p}(\hat{\boldsymbol{\beta}}_n) - \mathbf{p}(\boldsymbol{\beta}_n^*))'(\mathbf{p}(\hat{\boldsymbol{\beta}}_n) - \mathbf{p}(\boldsymbol{\beta}_n^*))}$$

  - The percentage of respondents for which the approach provides the smallest individual prediction error

Results

Scenario I

- 250 individuals participate in both the main experiment and the pilot study
- True choice behavior affected by one binary covariate z taking the values -1 or 1
  - $\Rightarrow$ Two covariate-based segments in the population with distinct mean choice behavior
- Heterogeneity distribution $\mathcal{N}(\beta_n^*|\boldsymbol{\Theta}^*\mathbf{z}_n, \boldsymbol{\Sigma}^*)$ with $\mathbf{z}_n = [1, z_n]'$ and

$$\boldsymbol{\Theta}^* = \begin{pmatrix} 0.5 & 1.5 \\ 0 & 0 \\ 0.5 & 1.5 \\ 0 & 0 \\ 0.5 & 1.5 \\ 0 & 0 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}^* = 0.5 \times \mathbf{I}_6$$

Results

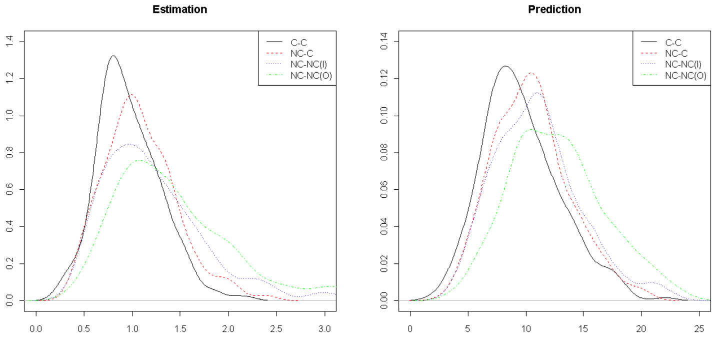
Scenario I

Estimation	*C-C*	*NC-C*	*NC-NC(I)*	*NC-NC(O)*
$RMSE_\beta$	**1.025**	1.150	1.340	1.579
Percentage (%)	**45.6**	16.4	22.8	15.2

Prediction	*C-C*	*NC-C*	*NC-NC(I)*	*NC-NC(O)*
$RMSE_p$	**10.208**	10.980	11.594	13.417
Percentage (%)	**42.8**	20.0	20.0	17.2

Results

Scenario I



Results

Scenario II

- 250 individuals participate in both the main experiment and the pilot study
- True choice behavior not affected by the covariate
  $\Rightarrow$ A single heterogeneous normal population
- Heterogeneity distribution $\mathcal{N}(\beta_n^*|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ with

$$\boldsymbol{\mu}^* = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}^* = 1.5 \times \mathbf{I}_6$$

Results

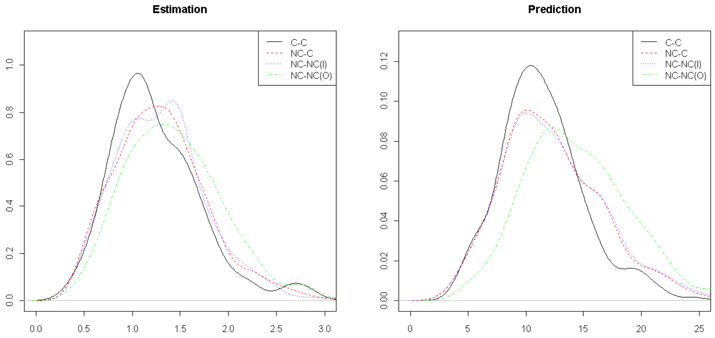
Scenario II

Estimation	*C-C*	*NC-C*	*NC-NC(I)*	*NC-NC(O)*
$RMSE_\beta$	**1.339**	1.393	1.380	1.532
Percentage (%)	**36.0**	14.8	25.6	23.6

Prediction	*C-C*	*NC-C*	*NC-NC(I)*	*NC-NC(O)*
$RMSE_p$	**12.101**	12.969	13.109	15.205
Percentage (%)	**43.6**	24.0	15.2	17.2

Results

Scenario II



Results

Scenario III

- 250 individuals participate in the main experiment, 100 additional persons are used in the pilot study
- True choice behavior affected by one binary covariate

Estimation	*C-C*	*NC-C*	*NC-NC(I)*	*NC-NC(O)*
$RMSE_\beta$	1.103	**1.100**	1.311	1.391
Percentage (%)	**35.6**	23.6	22.8	18.0

Prediction	*C-C*	*NC-C*	*NC-NC(I)*	*NC-NC(O)*
$RMSE_p$	**11.020**	11.503	12.307	13.067
Percentage (%)	**42.0**	24.4	18.4	15.2

Results

Scenario IV

- 250 individuals participate in the main experiment, 100 additional persons are used in the pilot study
- True choice behavior not affected by the covariate

Estimation	*C-C*	*NC-C*	*NC-NC(I)*	*NC-NC(O)*
$RMSE_\beta$	1.371	1.386	**1.359**	1.457
Percentage (%)	**35.6**	13.6	24.4	26.4

Prediction	*C-C*	*NC-C*	*NC-NC(I)*	*NC-NC(O)*
$RMSE_p$	**12.508**	12.870	12.967	14.662
Percentage (%)	**40.4**	16.4	18.8	24.4

Conclusions

This research shows the value of using covariates in individual experimental design and in hierarchical Bayes estimation of the mixed logit choice model

- When the covariates affect the true choice behavior of consumers, it is beneficial to include them in both individualized design and hierarchical Bayes estimation of the mixed logit choice model, C-C outperforms the other approaches
- In case respondents' true choice behavior is not impacted by covariates, either the C-C design and estimation approach remains superior or the decrease in estimation accuracy resulting from the inclusion of uninformative covariates is negligible
  - Only holds for a limited number of superfluous covariates
  - Do not add covariates indiscriminately but only use a few well thought variables