Peter Kurz & Stefan Binner

Which are the right Covariates in HB Estimation?

Introduction of Covariates

- Standard HB assumes one single multivariate normal population
- This could lead into a “shrinkage” of respondents into population mean
- HB could therefore reduce segment differences. The problem is greater with unbalanced proportional sample structure
- Significance testing of segment level differences is difficult (T-Test usually result in too small standard estimates within the segments)

➢ **Covariates in Upper Level Model** (CBC/HB)

- Respondents are no longer shrunk to one population mean, they are shrunk towards respondents with similar characteristics
- Only adds parameters for each different segment
- Covariance Matrix is shared for different prior distributions

Review of Recent Papers on Covariates

Peter Kurz & Stefan Binner: "Added Value through Covariates in HB Modeling?" *Sawtooth Software Conference 2010*

Keith Sentis and Valerie Geller: "The Impact of Covariates on HB Estimates" *Sawtooth Software Conference 2010*

Dimitri Liakhovitski & Faina Shmulyian: "Covariates in Discrete Choice Models: Are They Worth the Trouble?" ART-2011

Review of Recent Papers on Covariates

Peter Kurz & Stefan Binner: "Added Value through Covariates in HB Modeling?" *Sawtooth Software Conference 2010*

Basis for Analysis:

- 10 commercial studies
- 30.000+ interviews in total
- Covering industrial, durables, FMCG as well as B2B and B2C markets
- The studies were conducted in all parts of the world using state-of the-art computer aided interview delivery
- All studies developed in Sawtooth Software, analysis with HB

Review of Recent Papers on Covariates

Peter Kurz & Stefan Binner: "Added Value through Covariates in HB Modeling?" *Sawtooth Software Conference 2010*

Main Conclusions

- In most of our observations covariates in general did not improve results
- In some of our cases covariates distorted the priors, but converged to the same parameters
- In studies with large enough segment cells the covariate model converges towards same estimates as with standard HB (No influence of the covariate)
- If there is heterogeneity in the data which isn't assigned to the respondents multinormal distribution on attributes in best case we didn't improve results (sometimes we ended up with worse estimates)

Review of Recent Papers on Covariates

Peter Kurz & Stefan Binner: "Added Value through Covariates in HB Modeling?" *Sawtooth Software Conference 2010*

Further Hypothesis

- Covariates could improve results if we have clearly defined clusters with groups of respondents with different multinormal distributions on attributes
- One should first analyze the density of the common distribution of our data carefully and then decide whether or not we should use covariates or other techniques to improve the results.
- If there is a strong underlying cluster structure in the population, which was not taken into account in the sample planning and which can be identified and added as covariates to the HB estimation, this could help to improve the results.

Review of Recent Papers on Covariates

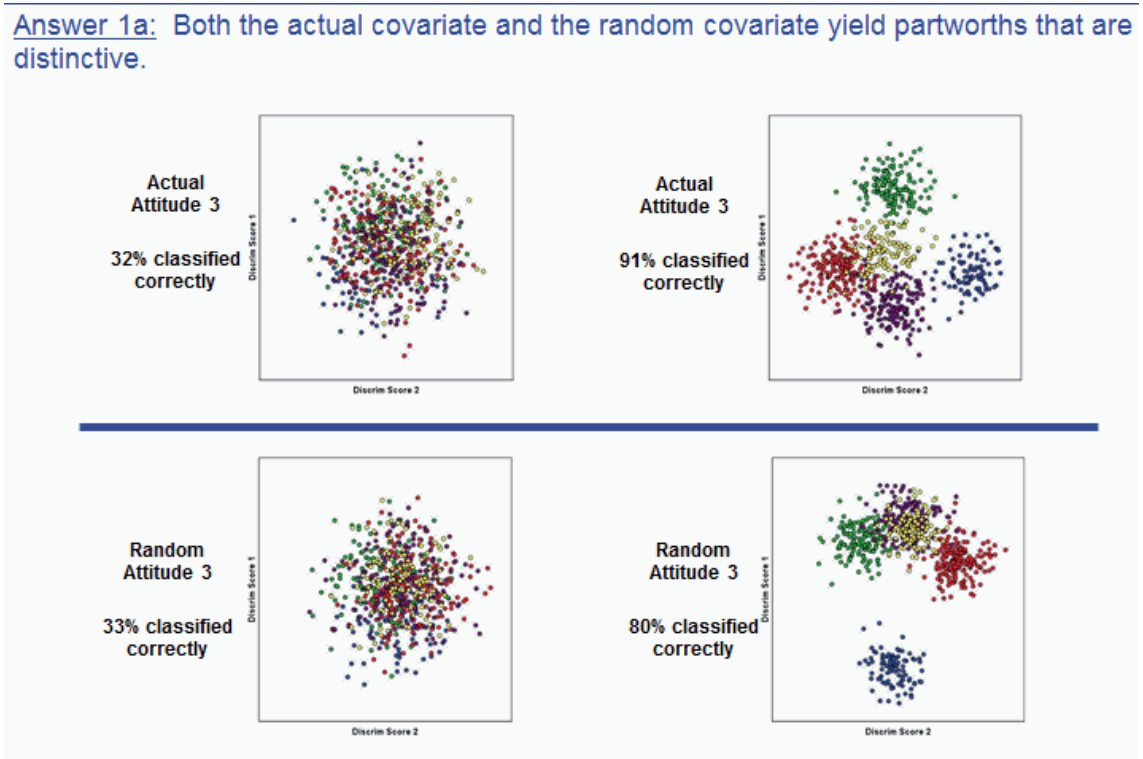
Keith Sentis and Valerie Geller: "The Impact of Covariates on HB Estimates" *Sawtooth Software Conference 2010*

Basis for Analysis:

- 5 datasets from commercial studies
- N = 8,445 Interviews
- Covering financial services, food, and beverages markets
- Total 40 attributes, 189 parameters, 43 covariates and 4,900,000 iterations
- All analysis with Sawtooth HB

Review of Recent Papers on Covariates

Keith Sentis and Valerie Geller: "The Impact of Covariates on HB Estimates" *Sawtooth Software Conference 2010*



Sentis and Geller were able to accurately recover the random segment membership using HB partworths with discriminance analysis .

Review of Recent Papers on Covariates

Keith Sentis and Valerie Geller: "The Impact of Covariates on HB Estimates" *Sawtooth Software Conference 2010*

Main Conclusions

- Across five datasets, fit and predictive accuracy are essentially unchanged by the inclusion of any of the 40+ covariates.
- Given the complete lack of impact, we failed in our quest to provide substantive advice about the three classes of covariates (Demographic, Category Behaviour, Attitudinal)
- HB partworths estimated using a random covariate are as specific to the segments' choice data as are the partworths estimated using the actual covariate

Review of Recent Papers on Covariates

Keith Sentis and Valerie Geller: "The Impact of Covariates on HB Estimates" *Sawtooth Software Conference 2010*

Further Hypothesis

- The HB part worths estimated using covariates are fitting "noise" rather than "signal" in the choice data of the different covariate segments
- Are there other hypotheses about why covariates do not increase predictive validity?

Review of Recent Papers on Covariates

Dimitri Liakhovitski & Faina Shmulyian: "Covariates in Discrete Choice Models: Are They Worth the Trouble?" ART-2011

Basis for Analysis:

- Synthetic data sets with up to N=160 (6 attributes with 4 levels)
- Generation of „true utilities" for the data sets
- Generation of „strong categorical covariates"
- HB Estimation using Sawtooth CBC HB and R package "bayesm"

Review of Recent Papers on Covariates

Dimitri Liakhovitski & Faina Shmulyian: "Covariates in Discrete Choice Models: Are They Worth the Trouble?" ART-2011

Main Conclusions

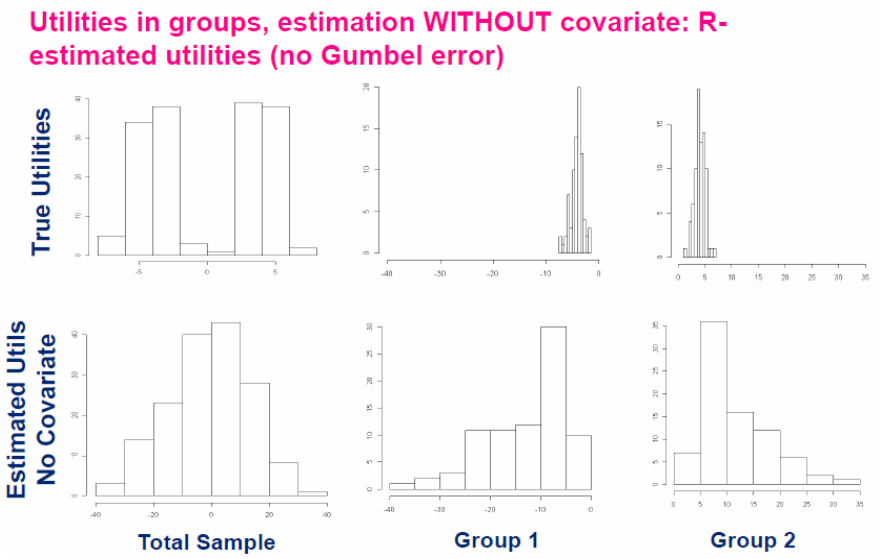
- For the data with known heterogeneous preference structure:
  - Using covariates in the HB estimation does not considerably improve the holdout hit rate and correlations with true utilities.
  - Even without a correct covariate, the HB estimation manages to converge to a sufficiently good solution
  - Sawtooth HB and R result in only minor differences in results
  - Sawtooth HB increases estimation time dramatically faster than R
- But:
  - HB estimation with covariates might result in better distributed utilities and in an improvement of aggregate metrics both at the total sample and subgroup level
  - Risk of introducing a "false" covariate does not seem high

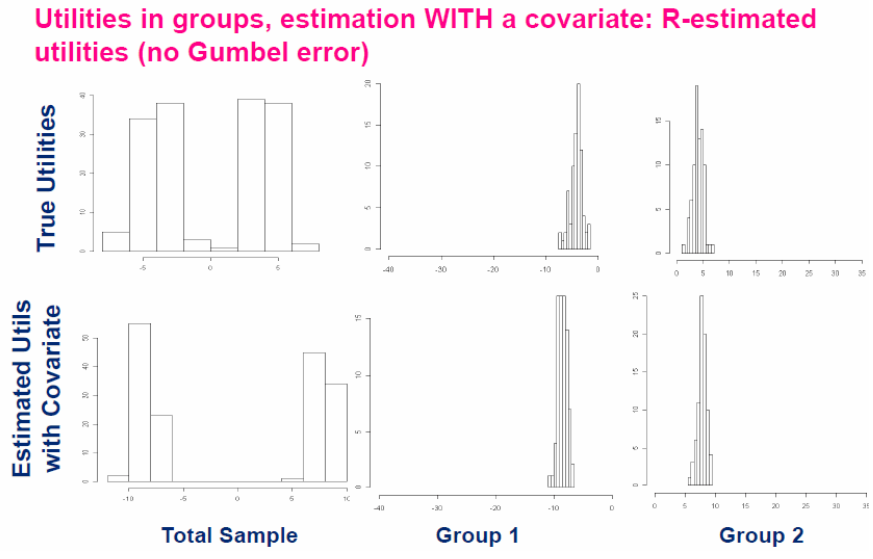
Review of Recent Papers on Covariates

Dimitri Liakhovitski & Faina Shmulyian: "Covariates in Discrete Choice Models: Are They Worth the Trouble?" ART-2011

Main Conclusions

Slide 25&26 ART Presentation Dimitri





HB estimation with covariates might result in better distributed utilities and in an improvement of aggregate metrics both at the total sample and subgroup level

Review of Recent Papers on Covariates

Dimitri Liakhovitski & Faina Shmulyian: "Covariates in Discrete Choice Models: Are They Worth the Trouble?" ART-2011

Further Hypothesis

- Limitations and Further Research Needs
  - Multiple covariates
  - Continuous covariates
  - Non "ideal" covariates
  - Partial covariates (those that exist only for some attributes)
  - Prior distribution and its parameters (tighter prior)
  - DCM layouts and designs
  - True utilities from actual studies
  - Ways to detect heterogeneity in samples - clustering

Strong Categorical Covariates

- Liakhovitski & Shmulyian used always a datasets with strong covariates and same number of levels for each attributes in dataset.

=>**What is a strong categorical covariate?**

Working Definition:
A covariate is “strong” if sub-groups defined by the covariate are very different with regard to their preferences (i.e. betas or utility parameters) and, at the same time, the members of each group are homogeneous with regard to their preferences

Need for a Complex Dataset

- "Strong categorical covariates" that impact all attributes are very rare in the reality
- Simulated datasets with realistic study designs (as can be found in daily research practice) are rare in academic papers
- To get more insight in the behavior of covariates we therefore need to simulate a more complex dataset with real structure!

Synthetic Data Set

- We rebuild one of our Datasets from the 2010 Sawtooth Conference Paper:
  - Car Study with 13 Attributes (1x5; 1x7; 11x3)
  - Number of Parameters 33
  - Proportional Sample: 6 Groups with N=500 Respondents each
  - 45.000 Choice Tasks
  - Design: 100 Versions using Sawtooth CBC/WEB 7
  - Answering behavior follows exactly the 6 different group characteristics (different buying intention)

design

interior

performance

safety

eco

price

Synthetic Data Set

- **Steps do create this dataset:**

  - ✓ Exporting CBC design to Excel
  - ✓ Calculating values for different purchase criteria (design, interior, performance, safety, eco, price)
  - ✓ Creating answers according to this six buying intentions
  - ✓ Sampling proportional dataset with 500 respondents for each of the 6 groups
  - ✓ Add 10% random error to the answers
  - ✓ HB Estimation with and without covariates
  - ✓ All HB settings default in Sawtooth Software CBC/HB 5.2

Example Choice Task:

- Showing that a concept could be optimal in more than one group

Which of these options would you choose?

Make	Audi	BMW	Mercedes-Benz	
Exterior design (according to the evaluated vehicles)	#	##	###	
Interior design (according to the evaluated vehicles)	###	##	#	
Driving performance	+ +	0	+	
Driving attributes	0	+ +	+	
Comfort	+ +	0	+	
Practibility	+ +	+	0	None of these options
Safety	+ +	0	+	
Quality	+	+ +	0	
Equipment level of the model series	+ +	0	+	
Sustainability (CO2, environmental standards)	0	+ +	+	
Economy	0	+	+ +	
Price in $	$40,000	$90,000	$100,000	
	○	○	○	○

Please, click on the bullet under the option you would choose.
And then click on the arrow symbol to continue.

- Improvement through covariates especially in such situations where respondents choice for one of the concepts could not be classified with 100% certainty.

Proportional Sample

- We can report significant differences in Shares, Hit-Rates, and MAE
- But especially in one group – the one we have shown with more then one optimal answer
- Random covariates results in same results than having no covariate
- Computational Time 10 to 20 times longer than without covariates

	hb prop	cov prop	ran prop
Design	0,29	0,00	0,33
Driving	0,11	0,00	0,07
Interior	4,14	0,01	4,34
Safety	2,58	0,02	2,64
Eco	0,44	0,00	0,45
Price	0,75	0,00	0,88
MAE	8,30	0,04	8,69
Rank	**2**	**1**	**3**

Disproportional Sample

- Same then in the disproportional case - we can show improvements in the problematic groups, but not in the disproportional segment.
- Estimates with covariate results in same amount of error than without.

Error	hp 65	cov 65	ran 65	hb 125	cov 125	ran 125	hb 250	cov 250	ran 250
Design	0,51	0,49	0,47	0,07	0,63	0,11	0,45	0,79	0,39
Driving	0,12	0,10	0,19	0,23	0,13	0,29	0,26	0,16	0,31
Interior	4,30	0,09	4,15	3,89	0,12	3,90	3,51	0,15	3,49
Safety	2,34	0,11	2,31	2,28	0,14	2,25	2,30	0,17	2,14
Eco	0,49	0,10	0,43	0,47	0,13	0,37	0,47	0,16	0,55
Price	0,85	0,10	0,77	0,86	0,13	0,91	0,95	0,16	0,89
MAE	8,61	1,00	8,33	7,80	1,28	7,82	7,93	1,60	7,78
Rank	**3**	**1**	**2**	**2**	**1**	**3**	**3**	**1**	**2**

Adding Error to the Covariate

- Adding a small amount of error (5%) to the covariate changed the results:

	hb prop	cov prop	cov prop 5%
Design	0,29	0,00	0,22
Driving	0,11	0,00	0,08
Interior	4,14	0,01	3,25
Safety	2,58	0,02	1,51
Eco	0,44	0,00	0,87
Price	0,75	0,00	1,18
MAE	8,30	0,04	7,12
Rank	**2**	**1**	**2**

- The positive effect of the covariate disappears immediately
- Here we see one reason why we can't show the positive effect of the covariates in real datasets.

CBC/HB vs. R

- We have re-run all the analysis done with Sawtooth CBC/HB Software again with the Bayesm package from Rossi et al.
- Differences only occur in case when having perfect data. R don‘t uses a dynamic update of step size to reach a certain acceptance rate. That’s the reason why R results have smaller variance than Sawtooth software.
- In all data set with a error component we have no significant differences in the results.
- R Code is much slower in computation than Sawtooth CBC/HB. Also on 64bit computers with up to 256 GB of Memory.
- The inclusion of covariates has nearly no influence on the computational time in R, but slow down Sawtooth CBC/HB by 10 to 20 times.

Summary

We could reproduce finding from all three previous papers:

- ✓ In data sets from real life studies, covariates normally do not influence share, hit rate and MAE
- ✓ Random covariates normally have no negative influence on share, hit rates and MEA, but influence the results of segmentation
- ✓ If covariates are chosen correct and goal of a study is segmentation, covariates could improve the quality of the segments
- ✓ In perfect situations (data and covariates without significant error) covariates improve share, hit rate, MAE and help to do better segmentation (academics are right that covariates improve estimates!)
- ✓ Using Sawtooth or Bayesm in R doesn't influence the results either with or without covariates.

Conclusions

- Covariates are a powerful extension to HB when you know exactly that the Data is structured by the covariate
- Covariates improve utilities especially when segmentation is your main goal
- But you have to be very careful when using covariates – small errors in the covariate or a covariate with different structure in data can harm your results
- Also segmentation can show miss leading results when the covariate has nothing to do with the choice data set
- If the error in the covariate is high or even random, the influence could sometimes be negative – but normally shares, hit rates and MAE are not influenced by the covariates

=> It depends on your data assumptions if covariates are useful or not for improving your estimates