A novel approach to detect attribute by covariate interactions in discrete choice models

Kyuseop Kwak, University of Technology Sydney, kyuseop.kwak@uts.edu.au
Paul Wang, University of Technology Sydney, paul.wang@uts.edu.au
Jordan Louviere, University of South Australia, jordan.louviere@unisa.edu.au

Abstract

This paper introduces a novel and simple method to identify attribute by covariate interactions in discrete choice models. This is important because incorporating such interactions in choice models can be an effective way to account for systematic taste variation or “observable preference heterogeneity” across individuals. Using simulated data sets to mimic a well-known phenomenon of selective attention to design attributes, we tested our proposed approach in a banking service context. Our proposed approach was successful in detecting the attribute by covariate interactions implied by the data generation process and outperformed a model with all covariate interactions. The proposed method contributes to the choice modelling literature by providing one of the “tricks of trade” to model observed preference heterogeneity. The simplicity of this approach has advantages for both academics and practitioners in marketing, transportation, healthcare and other fields that use choice modelling.

Keywords: Discrete choice model, attribute by covariate interaction

1. Introduction

Identifying and incorporating observable preference heterogeneity in discrete choice models applied to discrete choice experiments (DCEs, Louviere and Woodworth, 1983) remains a challenging issue. That is, one can account for systematic taste variation or observable preference heterogeneity across individuals by including attribute by covariate interactions in choice models (Louviere et al., 2000; Train, 2003). However, in practice this can be difficult to achieve because the number of covariates in DCEs often can be very large (e.g., 100 or more in online panel surveys). When one also considers that often one needs to dummy or effects code categorical covariates, this can lead to very large numbers of effects that one could consider. In turn, the larger the number of potential terms, the more observations are required to obtain reliable estimates; and a model with all possible or very many such interactions also may encounter multicollinearity associated with at least some of the interaction terms. To the best of our knowledge, there is little guidance in the literature on systematic identification and testing of attribute by covariate interactions. Thus, the purpose of this paper is to propose and describe a way to bridge this research gap by means of a novel method involving a new way to use relatively simple unconditional logit models (Long and Freese, 2006). We illustrate the method with both simulated and real discrete choice data sets.

2. Proposed method

A conditional logit (McFadden, 1974) is based on the assumption that individual i's utility of an alternative j in a choice task is a function of alternative j's attributes, i.e., x_{ij}. Thus, individual i's utility of alternative j is defined as:

(1) $$U_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \varepsilon_{ij}$$

Mixed (or random parameter) logit models that capture unobserved heterogeneity require one to assume that the marginal utilities for attributes, β, follow a particular distribution (Kamakura and Russell, 1989; Rossi and Allenby, 2003; Train, 2003). An alternative modelling approach is to specify a model with observed heterogeneity by including interactions between attribute x_{ij} and individual covariates, z_i as in (2) (individual covariates do not vary across choice tasks by an individual i). Parameters, γ capture such interaction effects, i.e.,

(2) $$U_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \gamma_1 x_{ij1} z_i + \gamma_2 x_{ij12} z_i + \cdots + \varepsilon_{ij}$$

Because DCEs can involve many covariates, the model specified in (2) can be extremely large and inefficient. To deal with cases that potentially involve many possible interactions, we propose a method that relies on unconditional logit (Long and Freese, 2006) or multinomial logit (Greene 2011)[1], which is easily implemented by many statistical software packages.

The proposed approach is as follows: 1) "Stack" the choice data such that each choice alternative represents a row of the data matrix. For example, if there are four choice alternatives in one choice task, there will be four rows for data from the DCE in the stacked dataset. 2) Filter the data by the chosen alternative (i.e., choice indicator = 1). That is, only "select" choice data for the chosen alternative for all subsequent analyses. 3) Specify an

[1] The term 'unconditional logit' is widely used in biostatistics and 'multinomial logit' is used in econometrics. Both terms refer to the same underlying model where the independent variables are individual specific characteristics that do not vary across choice alternatives.

unconditional logit model in which one attribute is the dependent variable, and all the observed individual characteristics (i.e., covariates) are independent or predictor variables. One needs to do this also for the Alternative Specific Constants (ASCs), if the design is a labelled DCE (Louviere et al., 2000). 4) The estimation results of the unconditional logit models identify potentially significant covariate interactions. 5) Finally, one tests the identified interaction effects by including them in a conditional logit model (McFadden, 1974).

The logic underlying this simple method is associated with the statistical analysis of contingency tables. Specifically, suppose one has a simple binary choice indicator (i.e., y = 0 or 1), a single binary attribute X coded as = -1 or 1, and a single binary covariate, Z coded as = -1 or 1. One can treat this as a contingency table where each cell represents the number of people who made a choice (i.e., y = 1) together with the binary attribute and binary covariate. For example, assume a DCE with 100 people, with preferences distributed as in Figure 1A, which represents no X and Z interaction, i.e., X and Y relationship do not vary due to Z, and as in Figure 1B which represents an interaction between X and Z.

		y	
Z	X	1	0
-1	-1	***25***	0
	1	0	***25***
1	-1	***25***	0
	1	0	***25***

Fig. 1A. NO interaction between X and Z

		y	
Z	X	1	0
-1	-1	***25***	0
	1	0	***25***
1	-1	0	***25***
	1	***25***	0

Fig. 1B. Interaction between X and Z

To illustrate our proposed approach, one first selects a column where y = 1. Then, one can reconfigure a contingency table with X and Z only as shown in Figures 2A and 2B. Now it should be obvious that Z will be significantly associated with X, which can be evaluated with a chi-square test. Figures 2A and 2B thus reveal that specifying an unconditional logit model with X as the dependent variable and Z as the independent variable should be able to test the statistical significance of Z. Although Figure 2B illustrates an extreme case of a perfect correlation, this approach also should work with less extreme correlations, as we now demonstrate.

	X	
Z	-1	1
-1	***25***	0
1	***25***	0

Fig. 2A. NO association between X and Z

	X	
Z	-1	1
-1	***25***	0
1	0	***25***

Fig. 2B. Significant association between X and Z

3. Monte-Carlo test of proposed approach

To test the generality and effectiveness of the proposed approach, we generated synthetic data sets based on a previous application by Kamakura et al. (1994), involving a banking choice experiment. To do this we created several DCE data sets with different sample sizes. We assumed two choice options (transaction account A or B) and four attributes: 1) minimum balance for fee waiver (MINBAL: $0, $500, $1000), 2) monthly check fee (CHECK: 0 cents, 15 cents, 30 cents), 3) monthly service fee (FEE: $0, $3, $6), 4) ATM options (ATM: n.a., free, 75 cents per use). We assumed that each person responded to nine choice sets.

We included observed heterogeneity (i.e., attribute by covariate interactions), based on the idea of selective attention to attributes (Bettman et al., 1991). To do this we created five covariates: Gender, Education, Income, Deposit Balance and Number of ATM Transactions (NATM). To make the data more realistic, we introduced correlations among some covariates, such as education and income with r = 0.7. Finally, we varied effect sizes (i.e., parameters magnitudes) of both main and interaction effects to simulate preference heterogeneity. The parameter setup is in Table 1; simulation codes and data are available from the authors as supplementary material.

Table 1. Parameter setup for Monte-Carlo DCE experiment

		Min. Bal	**(Min. Bal)^2**	**Check Fee**	**Mon. Fee**	**ATM fee**	**ATM 75c**
Main effect		*-0.6*	*0.18*	*-0.7*	*-0.3*	*0.5*	*-0.1*
Interactions	**Gender**	0	0	0	0	*-0.2*	*0.04*
	Education	*-0.1*	0	*-0.07*	*-0.01*	*0.2*	*-0.10*
	Income	*0.2*	0	*0.10*	*0.05*	0	*0.15*
	Deposit	*0.3*	0	*0.20*	*0.08*	0	0
	# of Account	0	0	*0.25*	0	*0.6*	*-0.20*

NOTE: Orthogonal polynomial coding is used for three quantitative attributes, i.e., minimum balance, check fee and monthly service fee. Quadratic main effects for check fee and monthly service fee are assumed to be zero. However, when we estimate a conditional logit model as presented in Table 3 (i.e., k = 8), both linear and non-linear (i.e., quadratic) main effects are included. For ATM fee, effects coding is used with 'ATM not available' as the reference category.

4. Test results

4.1 Detecting interactions using a series of unconditional logit model analyses

For each sample size, we estimated four unconditional logit models with each of the four design attributes as the dependent variable. The p-values from the unconditional logit results are in Table 2.

Table 2. p-values obtained from the unconditional logit analyses

Dependent variable = Minimum balance

covariates	n=100	n=300	n=600	n=900	n=1200
Gender (0)	0.900	0.422	0.329	0.220	**0.038**
Educ (-0.1)*	0.891	0.738	0.364	0.314	**0.009**
Income (0.2)	0.380	0.198	**0.005**	**0.000**	**0.000**
Deposit (0.3)	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**
Account (0)	0.344	0.747	*0.077*	*0.073*	**0.002**

Dependent variable = Check fee

covariates	n=100	n=300	n=600	n=900	n=1200
Gender (0)	0.283	0.168	0.222	0.515	0.166
Educ (-0.07)	0.924	0.537	0.264	0.409	0.625
Income (0.10)	0.797	0.497	0.181	0.169	0.289
Deposit (0.20)	0.658	**0.037**	**0.005**	**0.001**	**0.000**
Account (0.25)	**0.040**	**0.000**	**0.000**	**0.000**	**0.000**

Dependent variable = Monthly service fee

covariates	n=100	n=300	n=600	n=900	n=1200
Gender (0)	0.664	0.244	*0.071*	0.885	0.273
Educ (-0.01)	0.268	0.634	0.921	0.225	0.578
Income (0.05)	0.951	0.736	0.588	0.200	**0.044**
Deposit (0.08)	0.630	0.624	0.140	0.118	**0.036**
Account (0)	0.511	*0.086*	**0.001**	**0.000**	**0.003**

Dependent variable = ATM options

covariates	n=100	n=300	n=600	n=900	n=1200
Gender (0.04)	0.545	**0.001**	**0.000**	**0.000**	**0.000**
Educ (-0.10)	0.545	**0.003**	**0.007**	**0.000**	**0.000**
Income (0.15)	0.557	**0.002**	**0.005**	**0.001**	**0.000**
Deposit (0)	0.556	0.369	0.134	0.197	0.218
Account (-.20)	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**

* Values in parentheses are parameters assumed.

As expected, the results indicate that correct detection of interactions depends on sample size and parameter effect sizes. When effect sizes are small as in the case of monthly service fee, larger sample sizes are required. Nevertheless, even when effect sizes are moderate or large, the results suggest that the performance of our proposed approach with small sample size is as good as cases with larger sample sizes.

4.2 Specifying and running conditional logit model with interactions

We then estimated several conditional logit models, 1) main effects model, 2) main effects and the interactions identified in the previous step, i.e., proposed model, and 3) main effects and all possible interactions, i.e., full model. Fit statistics are in Table 3, which contains the results for three conditional logit model estimations. The results in Table 3 indicate that the proposed model significantly improves model fit compared to the main effects model and is parsimonious compared with the full model. The BIC criterion consistently picks the proposed model as the best model.

Table 3. Model fit statistics for conditional logit models

n=100	k	LL	AIC	BIC	ρ^2
Main Effects	8	-448.5	912.9	933.8	0.332
Proposed	12	-356.2	**736.4**	**767.7**	0.489
Full Model	48	-327.5	751.1	876.1	0.533

n=300	k	LL	AIC	BIC	ρ^2
Main Effects	8	-1379.9	2775.9	2805.5	0.318
Proposed	19	-1105.0	**2247.9**	**2318.3**	0.473
Full Model	48	-1077.5	2550.9	2428.7	0.484

n=600	k	LL	AIC	BIC	ρ^2
Main Effects	8	-2779.7	5575.5	5610.6	0.310
Proposed	21	-2185.1	4412.2	**4504.5**	0.482
Full Model	48	-2148.5	**4393.0**	4604.1	0.491

n=1200	k	LL	AIC	BIC	ρ^2
Main Effects	8	-5630.8	11277.6	11318.3	0.302
Proposed	26	-4382.2	**8816.3**	**8948.7**	0.479
Full Model	48	-4364.3	8824.7	9069.0	0.482

To compare biases in parameter recovery, we report Mean Absolute Percentage Error (MAPE) in Figure 3[2]. Across the different sample sizes, the results consistently indicate that a model based on the proposed approach produced smaller biases.

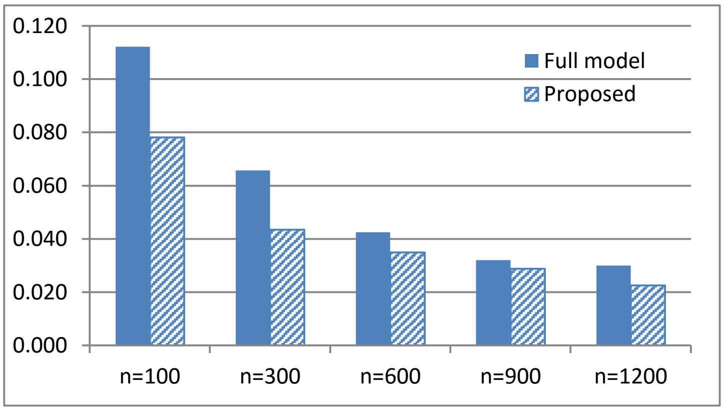

Fig. 3. Mean Absolute Percentage Error (MAPE) = $\left|\frac{(\beta_{true}-\beta_{estimate})}{\beta_{true}}\right|$

4.3 Applying the proposed approach to real DCE data

To illustrate the performance of our proposed approach, we applied it to two DCE choice data sets: a carbon trading scheme DCE and transport investment privatization DCE. The first

[2] Other measures such as Mean Error (ME), Mean Absolute Error (MAR) and Root Mean Squared Error (RMSE) also show similar pattern.

DCE, a carbon trading scheme, varies five 2-level attributes and 35 covariates that include consumer demographics, attitudes and opinions (More details can be found in Carson et al., 2010). Some covariates such as age groups and location are categorical variables with more than 10 categories. Even without dummy or effects coding, the number of possible interaction terms with five 2-level attributes and 35 covariates is 175. Despite being a seemingly small DCE, a model with all possible interactions is too large for a proper analysis. We used our proposed approach to identify 48 potentially significant interactions. We then tested the identified effects in a conditional logit model that specified the five main effects and the identified interactions. The associated fit statistics in Table 4 indicate a significant improvement relative to a main effects only model. Additionally, the identified interactions from our proposed approach provide useful insights in understanding heterogeneity in preferences for carbon trading schemes, as noted in Carson et al. (2010).

Table 4. Models with carbon trading scheme data

Model	k	LL	BIC	AIC	CAIC	ρ^2
Main Effects Only	5	-8219.50	16472.22	16449.00	16477.22	0.05
Main Effects plus Interactions	103	-7696.06	16076.43	15598.12	16179.43	0.13

The second DCE, transport privatization study, has three alternatives and nine attributes with 36 covariates. Using our proposed approach, we found 72 potentially significant interactions from 324 possible interactions. The final conditional logit provides very good fit statistics as shown in Table 5 and most interactions provide meaningful insights.

Table 5. Models with transport privatization study

Model	k	LL	BIC	AIC	CAIC	ρ^2
Main Effects only	8	-53746.37	107589.2	107508.7	107563.0	0.15
Main Effects plus Interactions	80	-52667.67	106300.1	105495.3	106038.0	0.17

5. Discussion and conclusions

We proposed a novel way to identify attribute by covariate interactions in discrete choice models. We used a synthetic data for a banking service DCE to test how well the approach recovers true effects. Specifically, we specified a particular data generation process with observed heterogeneity associated with five covariates. Overall, our proposed approach was successful in detecting the attribute by covariate interactions specified by the simulated data generation process. A potentially significant advantage of our proposed approach is that it does not require advanced statistical training and/or programming skills typically required to model unobserved heterogeneity with random coefficient or latent class choice models. Instead, our approach can be easily implemented with readily available commercial software packages such as SPSS, SAS or Stata.

Models that incorporate all possible attribute by covariate interactions are often difficult to implement in practice because of the large number of potential effects that must be estimated. For example, DCE surveys administered online often have more than 100 possible covariates as well as a large number of attributes. In such cases, if one tries to specify all possible attribute by covariate interactions in conditional logit models, one can easily exhaust the available degrees of freedom. This clearly suggests a need for a quick and easy way to identify covariates that interact significantly with ASCs and/or design attributes in DCEs.

Our proposed approach does not assume any prior knowledge regarding which covariates should be interacted with certain attributes. In many cases, researchers do not have a clear idea or hypotheses regarding socio-demographic covariates in discrete choice modelling. However, if someone has a good theoretical reason to believe that certain covariates should not affect respondents' choice decisions, those covariates should be removed from our procedure. Moreover, theory should help researchers specify the functional form of the choice model. Our approach is not considered a substitute for good theory underpinning the choice model. Also, our approach is not designed to test for higher level interaction between covariates such as gender and age interactions.

We view our results as a proof-of-concept, pilot test of the proposed approach; so it should be viewed as a starting point for further research. Of particular relevance would be a replication of our study in different research contexts using both stated and revealed preference data to determine the extent to which it generalizes. Further research should be conducted to assess statistical power of our approach. Another potentially worthwhile avenue of future research is using non-parametric statistical data mining methods like CART (Breiman et al., 1984) to detect localized interactions and incorporating the identified effects in conditional logit models to test their generality.

References

Bettman, J. R., Johnson E. J., Payne, J. W. 1991. Consumer decision making. In: Robertson, T. S., Kassarjain, H. H. (Eds), Handbook of Consumer Behaviour. Prentice-Hall, Englewood Cliffs, NJ, 50-84.

Breiman, L., Friedman, J. Olshen, R., Stone, C. 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Carson, R.T., Louviere, J.J. and Wei., E. 2010. Alternative Australian climate change plans: the public's views, Energy Policy 38 (2), 902-911.

Greene, William (2011), *Econometric Analysis 7th Edition*, Prentice Hall, New York.

Kamakura, W. A., Russell G. J. 1989. A probabilistic choice model for market segmentation and elasticity structure. Journal of Marketing Research 26 (4), 379-90.

Kamakura, W. A., Wedel, M., Agrawal, J. 1994. Concomitant variable latent class models for conjoint analysis. International Journal of Research in Marketing 11 (5), 451-464.

Louviere, J. J. and Woodworth, G. 1983. Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data, Journal of Marketing Research 20 (4), 350-367.

Louviere, J. J., Hensher, D. A., Swait J. 2000. Stated Choice Methods: Analysis and Application, Cambridge University Press, Cambridge.

Long, J. S., Freese, J., 2006. Regression Models for Categorical Dependent Variables Using Stats, 2nd Ed. Stata Press. College Station, Texas.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behaviour. Frontiers in Econometrics, ed. P. Zarembka, Academic Press, 105-42, New York.

Rossi, P., Allenby, G. 2003. Bayesian statistics and marketing. Marketing Science 22 (3), 304-328.

Train, K., 2003. Discrete Choice Methods with Simulation, Cambridge University Press, Cambridge, UK.