Detecting 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 Technology Sydney, jordan.louviere@uts.edu.au

Abstract

This paper introduces a simple way to identify attribute by covariate interactions in discrete choice models. This is important because modelling such interactions is an effective way to account for systematic taste variation or preference heterogeneity across different consumers. Using a simulated data set to mimic a well-known phenomenon of selective attention to design attributes, we tested our proposed approach in the banking service context. Our proposed approach was successful in detecting the attribute by covariate interactions implied by the data generation process and was found to outperform both full and stepwise interaction models. Such findings have implications for both academics and practitioners of the marketing research community in general and choice modelling field in particular.

Keywords: Discrete choice model, selective attention, covariate interaction

Detecting Attribute by Covariate Interactions in Discrete Choice Models

Introduction

In discrete choice experiment (DCE) models, product attributes generally vary over the choice alternatives (Street, Burgess and Louviere, 2005; Street and Burgess, 2007). Covariates, or characteristics of the decision-maker, on the other hand, do not vary over the alternatives. Theoretically speaking, an effective way to account for systematic taste variation or preference heterogeneity across different consumers is to incorporate attribute by covariate interactions into the choice model (Louviere, Hensher & Swait, 2000; Train, 2003). In practice, however, this modelling approach can be difficult to implement because of the large number of attributes and covariates involved.

Covariates may include socio-demographic variables such as gender and income, psychographic variables such as personality, and behavioural variables such as usage level. Apart from having multiple attributes for choice alternatives, a choice survey often includes many of these covariates. Finding meaningful interactions from all possible interactions can thus be a very challenging task. To our knowledge, there is little guidance in the literature on the systematic identification and testing of the attribute by covariate interactions. To bridge this research gap, we propose a novel method that involves a clever use of the relatively simple unconditional logit model (Long and Freese, 2006). We demonstrate this procedure using a simulated discrete choice data set.

The remainder of this paper is organized as follows. First, we briefly review previous literature regarding preference heterogeneity in choice data. Second, we introduce our proposed methodology to detect attribute by covariate interaction. Third, we discuss the results of testing our procedure using a simulated data set. Finally, we discuss the implications of our research findings and directions for further research.

Literature Review

In previous studies, unobserved heterogeneity is modelled in various ways. Most commonly, researchers make a certain distributional assumption for unknown heterogeneity across decision makers and across choice tasks. If a continuous distribution (e.g., multivariate normal distribution) is assumed, mixed logit (Train, 2003) or hierarchical Bayesian approaches (Rossi and Allenby, 2003) are used to approximate such unobserved heterogeneity. If a discrete distribution is assumed, finite mixture (or latent class) model is mostly used (Kamakura and Russell, 1989) to identify latent groups (i.e., segments) having same unobserved characteristics.

Although modelling unobserved heterogeneity with distributional assumption is statistically useful, such approaches pose a number of challenges to both marketing researchers and marketing managers. First, both mixed logit and latent class models require special training such as computer programming skills and high level of statistical knowledge (Train, 2003). Second, both modelling approaches require strong assumptions (Denzil et al., 2010) regarding distribution of model parameters or preference heterogeneity. Finally, preference heterogeneity uncovered from these methods often bears no or little relationship to observed

individual characteristics such as demographics or psychographic variables (Kamakura and Russell, 1989). As a result, managers find it hard to act upon findings from such modelling approaches.

Alternatively, covariates such as demographics and other consumer characteristics may explain individual heterogeneity. Individual covariates cannot be included as main effects in a conditional logit because covariates will not vary across choice tasks within each individual. Kamakura, Wedel & Agrawal (1994), for instance, extend the original latent class approach and incorporate covariates in identifying latent class membership of each individual. Another possibility is to include interactions between attributes and covariates in a typical DCE model (e.g., conditional logit).

By modelling observed covariates in a choice model, we can identify personal characteristics of individuals who focus on certain attributes only in decision making. However, this can be difficult to implement because the number of individual specific covariates in a choice experiment is sometimes very large (100 or more in online panels). Moreover, some individuals may not provide additional information to the choice decision process. Thus, we need a systematic approach to identifying and testing of the attribute by covariate interactions, which will be discussed in next section.

Methodology

Proposed Approach

We propose an approach based on simple and well-known unconditional logit (Long and Freese, 2006). One of the most important advantages is that it is easy to implement with commercial statistical software such as SPSS, SAS or STATA. The basic procedure is as follows. First, choice data needs to be stacked according to each choice alternative. For example, if there are four choice alternatives in one choice task, then there should be four rows for the choice task in the stacked dataset. Second, the stacked choice data are weighted using the dummy coded choice variable. If the option is chosen, then the choice variable takes the value of one. If the option is not chosen, the choice variable is set at zero. Third, the unconditional logit model is run using each attribute as a dependent variable, one at a time, and all the observed individual characteristics as the independent variables. Finally, significant independent variables are identified in the unconditional logit results. These variables are then included in the conditional logit model to be interacted with relevant attributes (McFadden, 1974).

Simulated Dataset

To test the effectiveness of our proposed approach, we generated a synthetic data set based on a previous application by Kamakura and Wedel (1994), where a banking choice experiment was conducted.

Following the optimal discrete choice design method (Street, Burgess and Louviere, 2005; Street and Burgess, 2007), we created a DCE with two options (transaction account A and B) using the following four attributes: 1) minimum balance for fee waiver (MINBAL: 0, $500, $1000), 2) monthly check fee (CHECK: 0, 15 cents, 30 cents), 3) monthly service fee (FEE: 0, $3, $6), 4) ATM options (ATM: n.a., free, 75 cents per use). The data was based on 1000

individuals and each individual received nine choice sets. We set marginal utilities for each of four attributes and assumed this applied to every individual.

We then created observed heterogeneity, i.e., attribute by covariate interactions, based on the idea of selective attention to attributes (Simone, 1955; Bettman, Johnson and Payne, 1991). We assumed three segments in the data set, where segment one pays attention to all four attributes, segment two pays attention to CHECK, FEE and ATM, and the final segment pays attention to two attributes: MINBAL and CHECK. Segment memberships are driven by five covariates: Gender, Education, Income, Deposit Balance, and Number of ATM Transactions (NATM). According to our setup, people with a low deposit balance are more likely to fall into segment two, those who pay less attention to minimum balance. People with higher education are more likely to be classified as segment three, those who do not care about service fee (FEE). People with higher income are classified as segment two and pay attention to ATM options more than people with lower income. Therefore, we created interactions between attributes and covariates in the simulated data set and expected our approach to detect these interactions.

Results

We followed our proposed procedure and ran four unconditional logit models. Table 1 suggests interactions which are consistent with data setup. Income was found to interact with all four attributes. Education was found to interact with MINBAL, FEE, and ATM. Monthly deposit only marginally interacts with MINBAL (p-value = 0.051). The other two covariates were found not to interact with any design attribute. Therefore, the approach suggests 10 interactions (due to separate coding for ATM_free and ATM_75cents) based on significant terms in Table 1.

Table 1 Unconditional Logit Results

	MINBAL		CHECK		FEE		ATM	
Predictors	Wald	p-value	Wald	p-value	Wald	p-value	Wald	p-value
EDUCATION	**22.941**	**0.000**	2.568	0.280	**20.434**	**0.000**	**19.680**	**0.000**
INCOME	**69.620**	**0.000**	**8.235**	**0.016**	**53.807**	**0.000**	**58.874**	**0.000**
DEPOSIT	**5.942***	**0.051**	1.185	0.550	1.322	0.520	1.370	0.500
NATM	1.967	0.370	0.121	0.940	1.880	0.390	1.695	0.430
GENDER	0.583	0.750	0.613	0.740	1.485	0.480	1.065	0.590

We then ran a conditional logit with five main effects plus 10 interactions. The fit statistics are presented in Table 2. Table 2 also shows two alternative conditional logit models for identifying attribute by covariate interactions. One is a full interaction model that contains five main effects plus all 25 possible interactions between five main effects and five covariates. The other approach is based on stepwise selection of significant interaction terms starting from full interaction model. It can be seen from Table 2 that our proposed model fits the data best in information criteria such as BIC and CAIC, thus offering strong empirical evidence in support of our proposed approach.

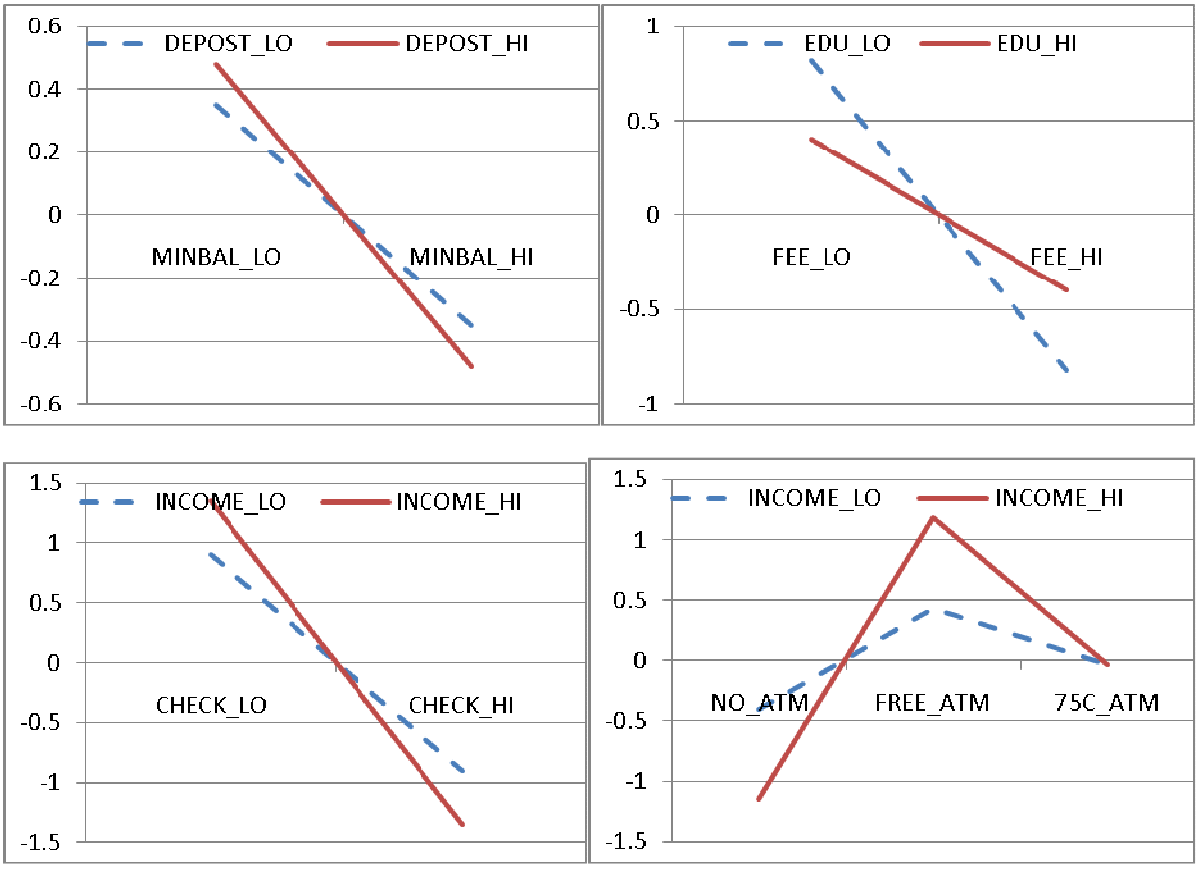
Table 2 Conditional Logit with Interactions

Model	k	LL	BIC	CAIC	R^2
Full Interaction	30	-3657.66	7522.56	7552.56	**0.462**
Stepwise Selection	17	-3668.83	7455.08	7462.99	0.461
Proposed	**15**	**-3672.19**	**7447.99**	**7462.08**	0.460

Note: k = number of parameters, LL = log-likelihood, BIC and CAIC are information criteria.

We have found interactions between four design attributes and five covariates to be statistically significant in our proposed model. For illustration purpose, we selected four interesting interactions shown in Figure 1. As expected, people with lower deposits are less sensitive to a minimum balance. People with higher income care less about monthly check fees, but they dislike no-ATM option and like free use of an ATM. In terms of education, highly educated people are less concerned about a monthly service fee than their less educated counterparts.

Figure 1 Selected Interaction Plots



Discussions and Conclusions

The paper has proposed a novel approach to identifying attribute by covariate interactions in a discrete choice modelling framework. In DCEs, we often observe the phenomenon of

selective attention to various attributes in a choice task. It is a well-documented phenomenon in consumer decision making literature. When faced with a large number of attribute and levels or profiles, consumers often use heuristic decision rules to decide which products to consider and choose. Consumers' selective attention to design attributes is an important source of preference heterogeneity. Using synthetic data in the banking services context, we modelled this data generation process and linked such heterogeneity to five observed covariates. Our proposed approach was successful in detecting the attribute by covariate interactions implied by the data generation process and was found to outperform both full and stepwise interaction models.

The full interaction models that incorporate all possible attribute by covariate interactions are often difficult to implement in practice because of the large number of potential covariates involved. With the advent of online panel providers, the number of possible covariates in a DCE can exceed 100. Given the need to dummy code categorical covariates such as location, household type, and occupation, one can quickly run out of degrees of freedom in the full interaction conditional logit models. The stepwise interaction conditional logit model can suffer from the same problem. Therefore, there is an urgent need for a quick and easy approach to identifying covariates that interact significantly with the design attributes in a DCE. To our knowledge, there is little guidance in the marketing literature on this topic and our proposed approach fills such a research gap.

As the first study of its kind, this paper provides a starting point for further research in this area. To enhance the generalisability of our research findings, it is worthwhile to replicate this study in various research contexts using both stated preference and revealed preference data in different product categories. Another avenue for future research is to try data mining non-parametric techniques such as CART (Breiman et al., 1984) or CHAID (Kass, 1980) to detect localized interactions.

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, pp. 50-84.

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

Denzil G. F., Keane M. P., Louviere, J. Wasi, N., 2010. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. Marketing Science 29 (3). 393-421.

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

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

Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29 (2), 119-127.

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.

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

Simone, H. A. 1955. A behavioural model of rational choice. Quarterly Journal of Economics 69, 99-118. .

Street, D. J., Burgess, L., Louviere, J. J. 2005. Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. International Journal of Research in Marketing 22 (4), 459-470.

Street, D. J., Burgess, L. 2007. The Construction of Optimal Stated Choice Experiments: Theory and Methods, Wiley, Hoboken, NJ.

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