A New Multiple Discrete-Continuous Choice Model for Percentage-Based Data

Xin Wang[1], Xiaoling Lu[1], and Hing-Po Lo[2]

[1]Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China

[2]Department of Civil Engineering, The University of Hong Kong, Hong Kong, China

Abstract

In recent years, the discrete-continuous models for multiple choices and allocation of time have seen great progress in consumer behavior analysis in marketing and transportation research. The models are based on the random utility maximization theory. They usually use Kuhn-Tucker conditions for likelihood maximization under non-negativity constraints and allow for both interior and corner solutions. However, to the best of our knowledge, all these models are applied to data in absolute scale. We know that quite often individuals allocate their total budget or total time on different commodities or activities in a percentage-based manner. In this paper, we propose a new multiple discrete-continuous choice model for this kind of data. The model formulation and estimation method are discussed, followed by simulation studies and real-life data analyses.

Key Words: Multiple discrete-continuous choice model; percentage-based data; random utility maximization theory

1. Introduction

Multiple discrete-continuous choice model is widely applied in time allocation and consumer consumption analysis. Wales and Woodland (1983) in the early eighties proposed a Kuhn-Tucker model to deal with customer behavior with corner solutions in simultaneous consumption, assuming that the utility function is maximized subject to the budget constraint and under a quadratic utility function. Kim et al. (2002) propose a new model based on the Kuhn-Tucker model, which is called KAR here. Their model is based on a translated, non-linear, but additive utility structure such that both interior and corner solutions are allowed, and with a diminishing marginal utility. The error term is assumed to have a normal distribution, and GHK method is used for high-dimensional integrals. Bhat (2005) instead uses an extreme value distribution in the error term. The likelihood function has a closed form and parameters can be easily estimated. Bhat calls his model the multiple discrete-continuous extreme value (MDCEV) model.

Besides that, Amemyia-Tobin model, which deals with share, is also proposed by Wales and Woodland. The model is based on Tobin (1958) and Amemyia (1974). Similarly, utility function is maximized subject to a budget constraint. A latent variable is assumed to explain the share. Also an error term is added to the latent variable. To ensure that share is between 0 and 1, truncated normal distribution is used. Dong et al. (2002,2003) use the Amemyia-Tobin model to analyze data. Based on Lee and Pitt (1986), Yen (2003) uses another form to deal with censored data. SML (Simulated Maximum likelihood Estimation) and QML (Quasi-maximum likelihood) are used to estimate parameters. All these models, complicated algorithms are needed to estimate parameters. In this paper, we propose a new multiple discrete-continuous choice model for share or percentage-based data. The rest of paper is organized as follows. I

2. The proposed model

2.1 Existing model for absolute scale data

Assuming there are K activities or commodities, Kim et al. (2002) and Bhat (2005) use the following utility function:

$$U = \sum_{k=1}^{K} \psi_k \left(x_k + \gamma_k \right)^{\alpha_k} e^{\varepsilon_k} \tag{1}$$

where $\psi_k, \alpha_k, \gamma_k$ are parameters and x_k is the quality of the kth commodity or time of the kth activity. The constraint for consumption problem is $\sum_{k=1}^{K} w_k x_k = E$, where w_k is the price for the kth commodity, and E is the total expenditure. The constraint for time allocation is $\sum_{k=1}^{K} x_k = T$, where T is the total time.

The two models are different in error term specification. In Kim et al. (2002), the error term was specified in the marginal utility. Assume that the deterministic marginal utility for kth alternative is $\bar{U}_k$, marginal utility with error term is U_k, then $\ln U_k = \ln \bar{U}_k + \varepsilon_k$. The distribution of error term is normal distribution. In Bhat (2005), the error term was specified in parameter ψ_k as follows: $\psi(x_k, \varepsilon_k) = \exp(\beta' y_k) \cdot \exp(\varepsilon_k)$, where y_k reflects the preference for alternative k. And the distribution of error term is extreme value distribution. When covariate variables are not considered, the two models are almost the same, except for the distribution of error term.

2.2 The new proposed model for percentage based data

Assuming the percentage value for kth alternative is p_k, then the utility for K alternatives is

$$U = \sum_{k=1}^{K} \frac{\psi_k}{\alpha_k + 1}\left[1 - (1 - p_k)^{\alpha_k + 1}\right] \tag{2}$$

where $\psi_k > 0,\ \alpha_k > 0$. In consumption problem $p_k = \frac{w_k x_k}{E}$, and in time allocation problem $p_k = \frac{x_k}{T}$. The utility for the kth activity is $\left[\psi_k / (\alpha_k + 1)\right] \cdot \left[1 - (1 - p_k)^{\alpha_k + 1}\right]$, which is an increasing function of p_k. And the marginal utility is, $\psi_k (1 - p_k)^{\alpha_k}$ which is also decreasing with p_k. All of these satisfy the definitions of utility and marginal utility in economics.

Assume that there are two activities, then $p_1 + p_2 = 1$. The slope of the indifference curve at any point is $slope(p_1, p_2) = -\frac{\partial U / \partial p_1}{\partial U / \partial p_2} = -\frac{\psi_1 (1 - p_1)^{\alpha_1}}{\psi_2 (1 - p_2)^{\alpha_2}}$. Different optimal results can be obtained based on different parameters as shown in Figure1.

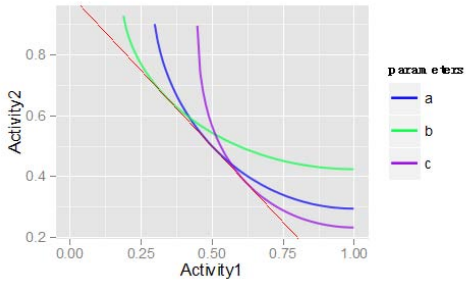

Figure1: Utility function and optimal percentage

$a: \psi_1 = \psi_2 = \alpha_1 = \alpha_2 = 1;\ b: \psi_1 = 1, \psi_2 = 2,\ \alpha_1 = \alpha_2 = 1;\ c: \psi_1 = \psi_2 = 1, \alpha_1 = 1, \alpha_2 = 2$

When α_k's are the same, the activity with larger ψ_k would have larger percentage value. When ψ_k's are the same, the activity with smaller α_k would have larger percentage value.

The marginal utility function of kth activity is $\psi_k (1 - p_k)^{\alpha_k}$, ψ_k can be interpreted as baseline marginal utility. When p_k is 0, the marginal utility is ψ_k, which is the largest marginal utility for activity k. For activity i and j, if ψ_i is larger than ψ_j, individual will obtain more utility by selecting activity i than j at the time point of 0. That is to say, individuals will be more likely to

select activity i. Overall, the number of corner solutions for activity i will be less than activity j. The role of α_k is to reduce the marginal utility when the percentage of activity k increases, which makes marginal utility satisfy the law of marginal diminishing. It can be called satiation parameter. Marginal utility will decrease quickly with larger α_k. If the percentage values of activities i and j are greater than 0, and α_i is greater than α_j, that means individual will have more utility if activity j has larger percentage value. Overall, the average percentage of activity j will be larger than that of activity i.

The utility function with error term is:

$$U = \sum_{k=1}^{K} \frac{\psi_k}{\alpha_k + 1}\left[1 - \left(1 - p_k\right)^{\alpha_k + 1}\right] \cdot e^{\varepsilon_k} \tag{3}$$

This model is called the multiple discrete-continuous percentage (MDCP) model. The Lagrangian function for the problem is:

$$L = \sum_{k=1}^{K} \frac{\psi_k}{\alpha_k + 1}\left[1 - \left(1 - p_k\right)^{\alpha_k + 1}\right] \cdot e^{\varepsilon_k} - \lambda\left[\sum_{k=1}^{K} p_k - 1\right] \tag{4}$$

The K-T conditions for optimal percentage p_k^* are given by:

$$\begin{aligned} &V_k + \varepsilon_k = V_1 + \varepsilon_1, \; if \;\; p_k^* > 0, \; (k = 2, 3, ..., K) \\ &V_k + \varepsilon_k < V_1 + \varepsilon_1, \; if \;\; p_k^* = 0, \; (k = 2, 3, ..., K) \end{aligned} \tag{5}$$

where $V_k = \ln \psi_k + \alpha_k \ln\left(1 - p_k\right)$. Set $d_i = \dfrac{a_i}{1 - p_i}$. If the distribution of the error term is extreme value distribution EV(0,1), the probability that M activities are selected from K activities is:

$$P\left(p_1^*, p_2^*, ..., p_M^*, 0, 0, ..., 0\right) = \left(\prod_{i=1}^{M} d_i\right)\left(\sum_{i=1}^{M} \frac{1}{d_i}\right)\left[\frac{\prod_{i=1}^{M} e^{V_i}}{\left(\sum_{j=1}^{K} e^{V_j}\right)^M}\right](M-1)! \tag{6}$$

The corresponding probability with EV(0, σ) as the error distribution is:

$$P\left(p_1^*, p_2^*, ..., p_M^*, 0, 0, ..., 0\right) = \frac{1}{\sigma^{M-1}}\left(\prod_{i=1}^{M} d_i\right)\left(\sum_{i=1}^{M} \frac{1}{d_i}\right)\left[\frac{\prod_{i=1}^{M} e^{V_i/\sigma}}{\left(\sum_{j=1}^{K} e^{V_j/\sigma}\right)^M}\right](M-1)! \tag{7}$$

If the normal distribution, N(0, Ω), is used in the error term, then the probability is

$$\begin{aligned} &P\left(p_1^*, p_2^*, ..., p_M^*, 0, 0, ..., 0\right) \\ &= \int_{-\infty}^{h_K} \cdots \int_{-\infty}^{h_{M+1}} \phi\left(h_2, ..., h_M, v_{M+1}, ..., v_K \mid 0, \Omega\right) abs\left|J\right| dv_{M+1}, ..., dv_K \end{aligned} \tag{8}$$

where $h_j = V_1 - V_j$, $v_j = \varepsilon_j - \varepsilon_1$, $j = 2, ..., M$, $\phi(\cdot)$ is the normal density, Ω is the covariance matrix, and J is the Jacobian.

In empirical analysis, covariate variables can be included in ψ_k and α_k as follows:

$$\ln\left(\psi_k\right) = \beta_{0k} + \beta_k' z + \varsigma y_k, \; k = 1, 2, ..., K \tag{9}$$

$$\ln\left(\alpha_k\right) = \eta_{0k} + \eta_k' z + \xi y_k, \; k = 1, 2, ..., K \tag{10}$$

where z represents individuals' attributes, and y_k represents variables related to alternative k, $\beta_{0k}, \beta_k, \varsigma, \eta_{0k}, \eta_k, \xi$ are parameters. When estimating parameters, $\psi_1 = 1$ if covariate variables are not considered, $\beta_{0k}, \beta_k, \varsigma, \eta_{0k}, \eta_k, \xi$ in ψ_1 are 0 if covariate variables are considered.

A drawback of the model is that it cannot deal with $p_k = 1$. When using this model, it requires that more than one alternative are selected by individuals.

3. Simulation Analysis

In this section, only compare the MDCP and MDCEV when applied to percentage-based data. There are three parts in this section. The first part is to use (3) to simulate data and use MDCP model and MDCEV model to estimate parameter. The second part is to use the utility function of the MDCEV model in Bhat (2005) to simulate data. The two models are used to estimate parameters. The third part is to use (3) to simulate 1000 different data sets, and their results are compared.

The parameters in the first part are $\psi=(1,3,4,2), \alpha=(5,2,3,4)$. The distribution of the error term is $EV(0,1)$. Data size is 1000. 750 initial values are used to estimate parameters. Here is the result:

Table1: Results for data based on the utility function of the MDCP model

		ψ_2	ψ_3	ψ_4	α_1	α_2	α_3	α_4	-logL
MDCP	Estimate	3.4018 1.5e-4	4.4995 1.9e-4	2.0797 8.6e-5	4.7831 1.5e-4	2.2604 5.4e-5	2.9965 1.0e-4	3.8880 1.1e-4	312.92 1.5e-5
	Standard error	0.2377 1.7e-5	0.3257 2.2e-5	0.1333 8.1e-6	0.2529 5.4e-6	0.0982 1.3e-6	0.1152 1.9e-6	0.1660 2.5e-6	
MDCEV	Estimate	1.5086 1.6132	1.5709 1.5943	1.2874 1.5116	0.0044 0.0033	0.0116 0.0166	0.0112 0.0116	0.0105 0.0126	2927.4 15.122

* Most results of MDCEV can't provide reasonable standard error. In each cell, figure above represents the average estimate of 750 initial values and the figure below is the standard deviation.

The results of MDCEV are not stable. But MDCP model can provide stable results with different initial values.

In the second part, $T \sim N(500,100)$, data size is 1000. The distribution of error term is $EV(0,1)$ EV$(0,1)$. Parameters are $\psi=(1,3,4,2), \alpha=(0.5,0.2,0.3,0.4)$. Cases with one activity selected are deleted. And the final data size is 958. For the data, percentage-based data can be found by t_k/T.

Table2: Results for percentage-based data based on MDCEV utility function

	ψ_2	ψ_3	ψ_4	α_1	α_2	α_3	α_4	ψ_2	-logL
MDCP	Estimate	0.9006 4.4e-5	1.0959 4.9e-5	0.9968 6.0e-5	1.3221 4.1e-5	5.6866 2.6e-4	1.8011 7.9e-5	1.3744 3.7e-5	669.96 2.8e-5
	Standard error	0.0468 3.5e-6	0.0587 4.4e-6	0.0532 5.1e-6	0.0556 1.3e-6	0.2040 7.9e-6	0.0713 2.2e-6	0.0593 1.1e-6	
MDCEV	Estimate	1.0081 1.1201	1.0761 1.2460	1.0600 1.0946	0.0063 0.0047	0.0093 0.0098	0.0107 0.0184	0.0101 0.0103	2904.9 15.636

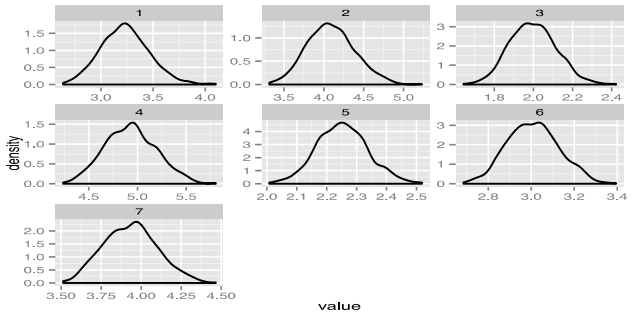

Figure 2: Density plot for each parameter

When data set based on utility function of MDCEV is transformed as percentage data, MDCEV can't provide stable results. However, MDCP can.

In the third part, 1000 different data sets are simulated based on the true parameters above for MDCP model in order to study the fluctuations of the estimates from the true parameters. Here is the result.

Plots 1-7 represent the density of estimate of $\psi_2, \psi_3, \psi_4, \alpha_1, \alpha_2, \alpha_3, \alpha_4$. Estimates based on different data sets are close to the true value without large bias. The density is close to normal distribution.

4. Empirical Data Analysis

The data source is based on a survey in Hong Kong in 2008. CATI is used to collected information on the use of time on different activities, and socio-demographics of Hong Kong residents. There are 412 successful interviews. Here pre-work or pre-school activities are considered. MAD (median absolute deviation, Hampel, 1985) is used to delete outliers. 396 cases are left. 209 are males 187 are females. 298 are workers and 98 are students.

There are almost ten activities included in the questionnaire. But some activities are selected only for few times. So the activities are regrouped into four activities, (1) Washing (W), (2) Dressing and make-up (D), (3) Breakfast (B), (4) Others (O). All respondents selected Washing, Dressing and make-up. 135 selected two activities. 176 selected three activities and 85 selected four. The following table shows some summary statistics of the four activities.

Table 3: Summary Statistics of the percentage of time used in the four activities

	All Data				Data with more than two selected			
Activity	W	D	B	O	W	D	B	O
Median	0.3798	0.2857	0.2566	0.3333	0.2857	0.2414	0.2566	0.3333
Mean	0.3922	0.3190	0.2983	0.3672	0.3067	0.2551	0.2983	0.3672
Sd	0.1936	0.1632	0.1668	0.1762	0.1513	0.1263	0.1668	0.1762
Frequency	396	396	184	162	261	261	184	162

There are two parts in the table, the first part is about the summary statistics of four activities for all data, and the second part is those for data with more than two selected. In the first part, the percentage of breakfast tends to be small and washing is large. It may be due to the fact that some respondents selected only two activities: washing, and dressing and make-up, which would make the percentage values of these two activities large. In the second part, the percentage of Others is the largest one. Dressing and make-up is the smallest.

MDCP model is used to analyze the time allocation data. Firstly, we use the model without covariate variables, and both normal distribution and extreme value distribution are used in the error term. And the maximum log-likelihood value is -262.39 and -225.26. So we use the extreme value distribution to build the model with covariate.

Table 4: Results of estimates without covariate

	ψ_2	ψ_3	ψ_4	α_1	α_2	α_3	α_4	-logL
EV(0,1)	1.2095	0.1659	0.1374	2.9024	4.4216	2.0730	1.3899	225.26
	(0.1537)	(0.0203)	(0.0172)	(0.1777)	(0.2454)	(0.2190)	(0.1753)	

Note: Figure above is the estimate and figure in parentheses is standard error.

The baseline marginal utility estimates of Washing ($\psi_1 = 1$), dressing and make-up (ψ_2) are close. And those of Breakfast (ψ_3) and Others (ψ_4) are close, which are smaller than the other two. The results are consistent with the data. The satiation parameter of Others (α_4) is the smallest, which indicates that if it was selected, the percentage would be large. And the satiation parameter of Dressing and make-up (α_2) is small. This means that the percentage of it would be small, which is consistent with the data.

Based on the model above, covariate variables are added in the model. Backward method is used to select variables.

Table 5: Results of estimates with covariate variables

	ψ_2	ψ_3	ψ_4	α_1	α_2	α_3	α_4
Intercept	0.0048	-1.8383*	-2.2755*	1.3824*	1.7724*	0.3163	0.3988
Female	-0.0247	0.0390	0.1602	-0.0266	-0.3382*	0.5742*	0.1275
25~34 years	-0.0655	-1.0559*	-0.3678	-0.1099	-0.1329	0.3241	-0.2338
35-44 years	0.4693	-0.0025	0.3185	-0.5574*	-0.2022	0.4387	-0.3206
More than 45 years	0.4141	0.5931	0.5525	-0.5001*	0.0374	0.2873	-0.1074

* Estimates are significant under 0.05 significant level.

From the above table, we can see that, satiation parameter of Washing, dressing and make-up of Female are smaller than that of male. It indicates that women tend to allocate more percentage of time in these activities. While, the behavior of having breakfast is different. Men tend to allocate more. People with 25~34 years are less likely to have breakfast. People with more than 35 years old are more likely to allocate more time for washing. The intercepts of ψ_3, ψ_4 are significant, which means that ψ_3, ψ_4 are significantly different from ψ_1.

5. Conclusions

In this paper we propose a new choice model for percentage based data. It satisfies the definitions of utility, marginal utility and the law of marginal diminishing. While, the model requires that more than one alternative are selected by individuals. In future research, we will improve the model to account for this situation.

Acknowledgement: This research is funded by a grant from the Ministry of Education, China under the Humanities and Social Sciences Project #11YJC910004. It is also supported by a grant from the Center for Applied Statistics, Renmin University of China.

Reference

Amemyia, T. (1974) "Multivariate regression and simultaneous equation models when the dependent variables are truncated normal," *Econometrica*, 42 (6), 999-1012.

Dong, D., B.W. Gould, and H.M. Kaiser (2002) "The structure of food demand in Mexico: an application of the Amemiya-Tobin approach to the estimation of a censored system," Selected paper in the 2002 Annual AAEA Meeting, Long Beach, CA.

Dong D. and H.M. Kaiser (2003) "Estimation of a censored AIDS model: a simulated Amemiya-Tobin approach," Cornell University, Department of Applied Economics and Management.

Bhat, C. (2005) "A multiple discrete–continuous extreme value model: formulation and application to discretionary time-use decisions," *Transportation Research Part B*, 39 (8), 679-707.

Bhat, C. (2008) "The multiple discrete-continuous extreme value (MDCEV) model:role of utility function parameters, identification considerations, and model extensions," *Transportation Research Part B*, 42 (3), 274-303.

Hampel, F.R. (1985). The Breakdown Points of the Mean Combined with Some Rejection Rules. Technometrics, 27 (2), 95-107.

Kim, J., G. M. Allenby and P. E. Rossi (2002) "Modeling consumer demand for variety," *Marketing Science*, 21 (3), 229-250.

Tobin, J. (1958) "Estimation of relationships for limited dependent variables," Econometrica, 26(1), 24-36.

Pollak, R.A. and T.J. Wales (1992), *Demand System Specification and Estimation*, Oxford University Press, New York.

Yen, S.T., B.H. Lin and D.M. Smallwood (2003) "Quasi- and simulated- likelihood approaches to censored demand systems: food consumption by food stamp recipients in the United States," *American Journal of Agricultural Economics*, 85 (2), 458–478.