Transportation Research Part B 63 (2014) 77–97

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

A multivariate hurdle count data model with an endogenous multiple discrete–continuous selection system

Chandra R. Bhat [a,b,*], Sebastian Astroza [a,1], Raghuprasad Sidharthan [c,2], Prerna C. Bhat [d,3]

[a] The University of Texas at Austin, Department of Civil, Architectural and Environmental Engineering, 301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, United States

[b] King Abdulaziz University, Jeddah 21589, Saudi Arabia

[c] Parsons Brinckerhoff, 999 3rd Ave, Suite 3200, Seattle, WA 98104, United States

[d] Harvard University, 1350 Massachusetts Avenue, Cambridge, MA 02138, United States

ARTICLE INFO

Article history:
Received 2 January 2014
Received in revised form 26 February 2014
Accepted 27 February 2014

Keywords:
Multivariate count data
Generalized ordered-response
Multiple discrete–continuous models
Hurdle model system
Endogeneity

ABSTRACT

This paper proposes a new econometric formulation and an associated estimation method for multivariate count data that are themselves observed conditional on a participation selection system that takes a multiple discrete–continuous model structure. This leads to a joint model system of a multivariate count and a multiple discrete–continuous selection system in a hurdle-type model. The model is applied to analyze the participation and time investment of households in out-of-home activities by activity purpose, along with the frequency of participation in each selected activity. The results suggest that the number of episodes of activities as well as the time investment in those activities may be more of a lifestyle- and lifecycle-driven choice than one related to the availability of opportunities for activity participation.

1. Introduction

In this paper, we develop a new econometric formulation and an associated estimation method for multivariate count data that are themselves observed based on a participation selection system. The participation selection system may be potentially endogenous to the multivariate count data in a hurdle-type model, which then leads to a joint count model system and participation selection system. The important feature of our proposed model is that the participation selection system itself takes a multiple discrete–continuous formulation in which multiple discrete states (with associated continuous intensities) may be simultaneously chosen for participation. A defining feature of our model is, therefore, that decision agents jointly choose one or more discrete alternatives *and* determine a continuous outcome as well as a count outcome for each discrete alternative. Further, if the decision agent does not choose a discrete alternative, there is no continuous or count outcome observed for this discrete alternative. Many empirical contexts in different fields conform to such a

* Corresponding author at: The University of Texas at Austin, Department of Civil, Architectural and Environmental Engineering, 301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, United States. Tel.: +1 512 471 4535; fax: +1 512 475 8744.

E-mail addresses: bhat@mail.utexas.edu (C.R. Bhat), sastroza@utexas.edu (S. Astroza), srprasad@utexas.edu (R. Sidharthan), prernabhat@college.harvard.edu (P.C. Bhat).

[1] Tel.: +1 512 471 4535; fax: +1 512 475 8744.

[2] Tel.: +1 206 382 5289; fax: +1 206 382 5222.

[3] Tel.: +1 512 289 0221.

http://dx.doi.org/10.1016/j.trb.2014.02.006
0191-2615/

decision framework and can benefit from our proposed model. For instance, consider an individual's daily engagement in non-work activities, an issue of substantial interest in the time-use and transportation fields. The individual chooses to participate in different activity types (such as shopping, visiting, and recreation), and jointly determines the amount of time to invest in each activity type and the number of episodes of each activity type to participate in. Of course, should an individual choose not to participate in a specific activity type, there is no issue of time investment and number of episodes associated with that activity type. Another example from the transportation and energy fields would be the case of a household's choice and use of motorized vehicles. Here, a household may choose to own different numbers of various body types of vehicles (such as a compact sedan and/or a pick-up truck), and put different mileages on the different vehicles. Again, the count and mileage are not relevant for body types not chosen by the household. Econometrically speaking, the potentially interrelated nature of the choices in these situations originates from common unobserved factors. For instance, underlying household factors such as environmental consciousness may make a household more likely to own multiple compact sedans and use compact sedans for much of the household's travel needs. These same unobserved factors can potentially also reduce the likelihood of the household owning one or more pick-ups and putting mileage on the pick-up(s).

Our formulation for the joint model combines a multiple discrete–continuous (MDC) model system with a multivariate count (MC) model system. The MDC system takes a MDC probit (MDCP) form in our formulation, while the MC system is quite general and takes the form of a multivariate generalized ordered-response probit (MGORP) model. In particular, we use Castro, Paleti, and Bhat's (CPB's) (2011) recasting of a univariate count model as a restricted version of a univariate GORP model. This GORP system provides flexibility to accommodate high or low probability masses for specific count outcomes without the need for cumbersome treatment (especially in multivariate settings) using zero-inflated mechanisms. The error terms in the underlying latent continuous variables of the univariate GORP-based count models for each discrete alternative also provide a convenient mechanism to tie the counts of different alternatives together in a multivariate framework. Further, these error terms form the basis for tying the MC model system with the MDCP model system using a comprehensive correlated latent variable structure. Overall, the model system extends extant models for count data with endogenous participation (for example, see Greene, 2009) that have focused on the simpler situation of a binary choice selection model and a corresponding univariate count outcome model.

The frequentist inference approach we use in the paper to estimate the joint MDCP-MC system is based on an analytic (as opposed to a simulation) approximation of the multivariate normal cumulative distribution (MVNCD) function. Bhat (2011) discusses this analytic approach, which is based on earlier works by Solow (1990) and Joe (1995). The approach involves only univariate and bivariate cumulative normal distribution function evaluations in the likelihood function (in addition to the evaluation of the closed-form multivariate normal density function).

The paper is structured as follows. The next section presents the modeling frameworks for the two individual components of the overall model system—the MDCP model and the MC model. This sets the stage for the joint model system formulated in this paper and presented in Section 3. Section 4 develops a simulation experiment design and evaluates the ability of the proposed estimation approach to recover the model parameters. Section 5 focuses on an illustrative application of the proposed model to the analysis of households' daily activity participation. Finally, Section 6 concludes the paper by summarizing the important findings and contributions of the study.

2. The individual model components

The use of the MDCP model in the current paper, rather than the MDC extreme value (MDCEV) model (Bhat, 2005, 2008) is motivated by the need to tie the MDC model with the MC model. For the MC model, as discussed in the previous section, we use a latent variable representation with normal error terms that also facilitates the tie with the MDCP model.

2.1. The MDCP model

Without loss of generality, we assume that the number of consumer goods in the choice set is the same across all consumers. Following Bhat (2008), consider a choice scenario where a consumer maximizes his/her utility subject to a binding budget constraint (for ease of exposition, we suppress the index for consumers):

$$\begin{aligned} \max \quad & U(\mathbf{x}) = \sum_{k=1}^{K} \frac{\gamma_k}{\alpha_k} \psi_k \left(\left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right) \\ s.t. \quad & \sum_{k=1}^{K} p_k x_k = E, \end{aligned} \tag{1}$$

where the utility function $U(\mathbf{x})$ is quasi-concave, increasing and continuously differentiable, $\mathbf{x} \geqslant 0$ is the consumption quantity (vector of dimension $K \times 1$ with elements x_k), and γ_k, α_k, and ψ_k are parameters associated with good k. In the linear budget constraint, E is the total expenditure (or income) of the consumer ($E > 0$), and p_k is the unit price of good k as experienced by the consumer. The utility function form in Eq. (1) assumes that there is no essential outside good, so that corner solutions (*i.e.*, zero consumptions) are allowed for all the goods k (though at least one of the goods has to be consumed, given a positive E). The assumption of the absence of an essential outside good is being made only to streamline the presentation; relaxing