Transportation Research Part B 91 (2016) 52–76

Contents lists available at ScienceDirect

Transportation Research Part B

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

Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: Application to residential self-selection effects analysis in an activity time-use behavior model

Chandra R. Bhat[a,b,*], Sebastian Astroza[a,1], Aarti C. Bhat[c], Kai Nagel[d,2]

[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] The University of Texas at Austin, College of Natural Sciences and Liberal Arts, Austin, TX, United States
[d] Technische Universität Berlin, Transport Systems Planning and Transport Telematics, Sekr. SG12, Salzufer 17-19, 10587 Berlin, Germany

ARTICLE INFO

Article history:
Received 28 August 2015
Revised 11 March 2016
Accepted 13 March 2016
Available online 14 May 2016

Keywords:
Latent factors
Mixed dependent variables
Structural equations models
MACML estimation approach
Residential self-selection effect
Activity time-use

ABSTRACT

This paper makes both a methodological contribution as well as an empirical contribution. From a methodological perspective, we propose a new econometric approach for the estimation of joint mixed models that include a multiple discrete choice outcome and a nominal discrete outcome, in addition to the count, binary/ordinal outcomes, and continuous outcomes considered in traditional structural equation models. These outcomes are modeled together by specifying latent underlying unobserved individual lifestyle, personality, and attitudinal factors that impact the many outcomes, and generate the jointness among the outcomes. From an empirical perspective, we analyze residential location choice, household vehicle ownership choice, as well as time-use choices, and investigate the extent of association versus causality in the effects of residential density on activity participation and mobility choices. The sample for the empirical application is drawn from a travel survey conducted in the Puget Sound Region in 2014. The results show that residential density effects on activity participation and motorized auto ownership are both associative as well as causal, emphasizing that accounting for residential self-selection effects are not simply esoteric econometric pursuits, but can have important implications for land-use policy measures that focus on neo-urbanist design.

1. Introduction

The joint modeling of multiple outcomes is of substantial interest in several fields. In econometric terminology, this jointness may arise because of the impact (on the multiple choice outcomes) of common underlying exogenous observed variables, or common underlying exogenous unobserved variables, or a combination of the two. For instance, consider the

* 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), aartibhat@utexas.edu (A.C. Bhat), nagel@vsp.tu-berlin.de (K. Nagel).
[1] Tel.: +1 512 471 4535; fax: +1 512 475 8744.
[2] Tel.: +49 30 314 23308; fax: +49 30 314 26269.

http://dx.doi.org/10.1016/j.trb.2016.03.007

choice of residential location, motorized vehicle ownership (or simply auto ownership from hereon), and activity time-use in recreational pursuits (such as going to the movies/opera, going to the gym, playing sports, and camping). In this setting, it is possible (if not very likely) that individuals from households who have a high green lifestyle propensity (an unobserved variable) may search for locations that are relatively dense (with good non-motorized and public transportation facilities and high accessibility to activity locations), may own fewer cars, may travel less and so pursue more in-home (IH) activities, and pursue less of what they may perceive as activities that correlate with extravagant living and indulgence such as out-of-home (OH) personal care/grooming, shopping, and dining out. In this case, when one or more unobserved factors (for example, green lifestyle) affect(s) the multiple outcomes, independently modeling the outcomes results in the inefficient estimation of covariate effects for each outcome (because such an approach fails to borrow information on other outcomes; see Teixeira-Pinto and Harezlak, 2013). But, more importantly, if some of the endogenous outcomes are used to explain other endogenous outcomes (such as examining the effect of density of residence on auto ownership, or the effect of density of residence on OH activity time-use, or the effect of auto ownership on time-use in activities), and if the outcomes are not modeled jointly in the presence of unobserved exogenous variable effects, the result is inconsistent estimation of the effects of one endogenous outcome on another (see Bhat and Guo, 2007, and Mokhtarian and Cao, 2008). In the next section, we position the current paper within this broader methodological context of modeling multiple outcomes jointly.

1.1. The methodological context

The joint modeling of multiple outcomes has been a subject of interest for many years, dominated by the joint modeling of multiple continuous outcomes (see de Leon and Chough, 2013). However, in many cases, the outcomes of interest are not all continuous, and will be non-commensurate (that is, a mix of continuous, count, and discrete variables). The joint modeling of non-commensurate outcomes makes things more difficult because of the absence of a convenient multivariate distribution to jointly (and directly) represent the relationship between discrete and continuous outcomes. This is particularly the case when one of the dependent outcomes is of a multiple discrete-continuous (MDC) nature. An outcome is said to be of the MDC type if it exists in multiple states that can be jointly consumed to different continuous amounts. In the example presented in the earlier paragraph, activity time-use is an MDC variable, assuming a daily or weekly or monthly period of observation. Thus, in a given day, an individual may participate in multiple types of non-work activities (shopping, personal business, child-care, recreation, and so on) and invest different amounts of time in each activity types (see Bhat et al., 2009 and Pinjari and Bhat, 2014 for detailed reviews of MDC contexts).

In this paper, we introduce a joint mixed model that includes an MDC outcome and a nominal discrete outcome, in addition to count, ordinal, and continuous outcomes. Each non-continuous outcome is cast in the form of a latent underlying variable regression, wherein the latent "dependent" stochastic variable is assumed to manifest itself through an a priori transformation rule in the observed non-continuous outcomes. Next, the continuous observed outcome and the latent continuous manifestations of the non-continuous dependent outcomes themselves are tied together using a second layer of common latent underlying unobserved decision-maker variables (such as individual lifestyle, personality, and attitudinal factors) that impact the outcomes. The presence of this second layer of latent "independent" is what generates jointness among the outcomes. Reported subjective ordinal attitudinal indicators for the latent "independent" variables help provide additional information and stability to the model system. In this manner, we build on Bhat's (2015) Generalized Heterogeneous Data Model (GHDM) that expressly acknowledges the presence of latent "independent" variables (or sometimes referred to as latent psychological constructs in the social sciences and in this paper as well) affecting choice, and assumes that these latent "independent" variables get manifested in observed psychological indicators as well as the observed dependent outcomes. In particular, we develop a powerful and parsimonious way of jointly analyzing mixed outcomes including an MDC outcome. In addition, we formulate and implement a practical estimation approach for the resulting GHDM (GHDM including an MDC outcome) model using Bhat's (2011) maximum approximate composite marginal likelihood (MACML) inference approach. This approach is not simulation-based (see Bhat, 2000 and Bhat, 2001 for such simulation approaches, but which can lead to convergence issues as well as be computationally intensive). Rather, the MACML approach requires only the evaluation of bivariate or univariate cumulative normal distribution functions regardless of the number of latent variables or the number and type of dependent variable outcomes. Many structural equation models (SEMs) and similar models in the past, on the other hand, are estimated using simulation-based methods or, alternatively, sequential estimation methods (see Temme et al., 2008 and Katsikatsou et al., 2012 for discussions of these sequential methods). The problem with the latter sequential methods is that they do not account for sampling variability induced in earlier steps in the later steps, leading to inefficient estimation. In addition, the use of such sequential methods will, in general, also lead to inconsistent estimation (see Daziano and Bolduc, 2013 for discussions of the reasons). The MACML approach is a practical way to obtain consistent estimators even in high dimensional mixed multivariate model systems.

To our knowledge, this is the first formulation and application of such an integrated model system in the econometric and statistical literature. The model should be applicable in a wide variety of fields where MDC variables appear as elements of package choices of different types of outcomes of interest. For example, in the health field, in addition to binary, count, and continuous variables related to the occurrence, frequency, and intensity, respectively, of specific health problems, it is not uncommon to obtain ordinal information on quality of life outcomes/perceptions and there may be interest in associating these variables with an MDC variable representing the type and intensity of participation in different types of physical activities and the durations in each participated physical activity. Other fields where the proposed model should be of