ARTICLE IN PRESS

Journal of Choice Modelling xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Choice Modelling

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

Analyzing the continuity of attitudinal and perceptual indicators in hybrid choice models

Francisco J. Bahamonde-Birke[a,b,c,*], Juan de Dios Ortúzar[d]

[a] Institut für Verkehrsforschung, Deutsches Zentrum für Luft-und Raumfahrt (DLR), Germany
[b] Energy, Transportation and Environment Department, Deutsches Institut für Wirtschaftsforschung, Berlin, Germany
[c] Technische Universität Berlin, Germany
[d] Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Chile

ARTICLE INFO

Keywords:
Latent variables
Hybrid discrete choice models
Attitudinal and perceptual indicators
Continuity

ABSTRACT

The main objective of this paper is to compare the consequences of treating the attitudinal and perceptual indicators of hybrid discrete choice (HDC) models as continuous or ordinal outcomes. Based on tradition and computational reasons, such indicators are still predominantly treated as continuous outcomes in practice. This usually neglects their nature (as respondents are normally asked to state their preferences, or level of agreement with a set of statements, using a discrete scale) and may induce bias.

We conducted an analysis based on simulated data and real information (two case studies) and were able to find that the distribution of the indicators (especially when associated with non-uniformly spaced thresholds) may lead to a clear deterioration of the model's overall predictive capacity, when assuming continuous indicators. This, however, does not translate to the goodness-of-fit of the discrete-choice component of the HDC model. Along the same line, a higher relative variability of the latent variables increases the differences between both approaches (ordinal and continuous outcomes), especially concerning the goodness-of-fit of the discrete-choice component of the model. It was not possible to identify a relation between the predictive capacity of both approaches and the amount of available information. Considering more indicators tend to reduce the gap between both approaches, but the effect is significantly smaller than the effect of the relative variability.

Finally, two case studies using real data confirmed that no major differences, regarding the predictability of the discrete choices can be observed when the relative variability of the latent variables is low. Thus, we recommend analyzing this relative variability in order to decide on the suitability of the continuous assumption as a suitable alternative to the more onerous but correct ordinal treatment.

1. Introduction

The effort of enriching discrete choice (DC) models (McFadden, 1974) with unobserved latent constructs, accounting for unobserved factors that may be relevant in the decision making process, led to the first versions of the so called hybrid discrete choice (HDC) models (McFadden, 1986; Train et al., 1987) during the 80s. Despite not having a great impact in its origins (mainly due to computational issues), HDC modelling was revitalized during the last decade (Ben-Akiva et al., 2002). Since then, the

* Corresponding author at: Institut für Verkehrsforschung, Deutsches Zentrum für Luft-und Raumfahrt (DLR), Germany.
E-mail addresses: bahamondebirke@gmail.com (F.J. Bahamonde-Birke), jos@ing.puc.cl (J.d.D. Ortúzar).

http://dx.doi.org/10.1016/j.jocm.2017.01.003
Received 19 January 2017; Accepted 21 January 2017
1755-5345/

Please cite this article as: Bahamonde-Birke, F.J., Journal of Choice Modelling (2017), http://dx.doi.org/10.1016/j.jocm.2017.01.003

approach has gained popularity and has become a standard tool in travel behaviour research (van Acker et al., 2011; Ashok et al., 2002; Raveau et al., 2012; Alvarez-Daziano and Bolduc, 2013; Bahamonde-Birke and Ortúzar, 2014a; among many others).

Basically, HDC models are regular DC models that include latent constructs as explanatory variables. These so called *latent variables* (LV) aim to capture unobservable/unobserved attributes of individuals and alternatives, such as attitudes, perceptions, values, etc. Thus, their inclusion into the DC-framework allows to analyze how these underlying factors affect the decision making process, increasing the predictive capability and realism of econometric models. As LV cannot be directly measured by the modeller they are usually estimated relying on a *Multiple Indicators MultIple Causes* (MIMIC) model (Zellner, 1970; Bollen, 1989). Here, it is assumed that the LV are explained by a set of characteristics of the users and the alternatives (through so called *structural equations*), while explaining, at the same time, a set of indicators previously gathered from the individuals (through so called *measurement equations*).

The aforementioned indicators play a key role in the identification of latent variables. They are supposed to be a measurable expression of underlying unobservable/unobserved characteristics (to be modelled via LV) and are normally gathered by asking respondents to state their preferences, or level of agreement, with a set of statements using a discrete scale (Likert, 1932).

For the sake of simplicity as well as for historical reasons (the first models were based on that assumption; Morikawa et al., 1996; Ben-Akiva et al., 2002), indicators are usually considered a linear continuous expression of the LV in the context of HDC models (Vredin-Johansson et al., 2006; Yáñez et al., 2010; Alvarez-Daziano and Barla, 2012; Bahamonde-Birke and Ortúzar, 2014b, among many others). This assumption, however, neglects the nature of the indicators (when collected using a Likert scale[1]) and has the potential to induce bias in the estimation. In fact, it has been argued that the indicators should not be treated as discrete but rather as ordinal outcomes (Daly et al., 2012).

Even though an increasing number of researches has started to rely on ordinal assumptions (e.g. Czajkowski et al., 2015; Kamargianni et al., 2015; Mariel and Meyerhoff, 2016), a significant amount of the literature is still based on the assumption of continuity (e.g. Motoaki and Daziano, 2015; Bahamonde-Birke and Hanappi, 2016; Fernández-Heredia et al., 2016; Schmid et al., 2016; Valeri and Cherchi, 2016). This phenomenon should not only be attributed to historical reasons but also to practical issues, as considering the indicators as ordinal outcomes has major implications in terms of computational costs. In particular, model estimation gets considerable more involved, especially when considering specifications not leading to close-form expressions for the probabilities, such as the *Ordered Probit* model (OPM). That is the reason why researchers considering ordinal indicators, tend to rely on the *Ordered Logit* model (OLM; Daly et al., 2012; Hess et al., 2013).

This paper analyzes the performance of both modelling approaches (treating the indicators continuously and in an ordered fashion) under different situations; in particular we consider the distribution of the indicators, their variances, the sample size and the number of levels posed in the Likert scales, to establish under which conditions the continuous treatment may be considered a suitable alternative. The rest of the paper is organized as follows. Section 2 offers a theoretical overview of the modelling background. Section 3 presents an analysis based on simulated data (considering several Monte Carlo simulations), while Section 4 deals with the implication of the continuity assumptions on two real data case studies. Finally, Section 5 summarizes the paper's conclusions.

2. Theoretical background

As mentioned in the previous section, latent variables (LV) are normally considered in order to include unobserved factors, such as attitudes, perceptions, values, etc., into the modelling. For their estimation, the analyst usually relies on MIMIC models assuming the existence of LV, which are a function of positively observed explanatory variables and, eventually, of other latent constructs (Kamargianni et al., 2014; Link, 2015). This way, the above-mentioned *structural equations* may take the following form (assuming a linear specification):

$$\eta = X\cdot\alpha_X + \eta^{*}\cdot\alpha_\eta + \upsilon \tag{2.1}$$

Here, η and η^* are vectors describing sets of jointly dependent endogenous latent constructs, X is a set of exogenous observed explanatory variables and υ an error term that can follow any distribution, but is usually assumed to be Normal with mean zero and a given covariance matrix Σ_η. α_x and α_η are vectors of parameters to be estimated.

In a MIMIC structure, this set of equations will always be unidentified so it is necessary to consider it in conjunction with a *measurement equations* set. The latter may be described in the following manner (again assuming a linear specification):

$$I = X\cdot\gamma_X + \eta\cdot\gamma_\eta + \varsigma \tag{2.2}$$

where I is a vector of exogenous indicators and ς an error term, the distribution of which will depend on the assumptions regarding the indicators. Finally γ_x and γ_η are vectors of parameters to be estimated.

An advantage of this specification is that once the MIMIC model has been estimated, the LV are independent from the indicators (which are normally neither useful for forecasting nor for explanatory purposes), as they can be estimated directly from the structural equations. As a consequence, the LV are intrinsically associated with the explanatory variables considered in these equations (which are typically socioeconomic variables describing the individuals) and are a depiction of the situation when the

[1] Even if the modeller would allow for individuals to state continuous indicators, it is highly doubtful whether respondents would take the whole scale into consideration and decimal numbers may be underrepresented in favor of integers. Even more, it is a debatable point whether respondents consider all values in a Likert-scale equally, as simplifying heuristics (Tversky and Kahneman, 1974) may cause some levels to be ignored (leading for instance to an overrepresentation of the extremes and the midpoint).