Accident Analysis and Prevention 55 (2013) 1–11

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Evaluating alternate discrete choice frameworks for modeling ordinal discrete variables

Naveen Eluru*

Department of Civil Engineering and Applied Mechanics, McGill University, Suite 483, 817 Sherbrooke St. W., Montréal, Québec H3A 2K6, Canada

ARTICLE INFO

Article history:
Received 28 August 2012
Received in revised form 5 January 2013
Accepted 8 February 2013

Keywords:
Ordered and unordered discrete choice models for injury severity
Ordinal discrete variables
Generalized ordered logit
Comparison

ABSTRACT

There is considerable debate on the appropriate discrete choice framework for examining injury severity. Researchers in the safety field have employed ordered and unordered frameworks for examining the various factors influencing injury severity. The objective of the current study is to investigate the performance of the ordered and unordered response frameworks at a fundamental level. Towards this end, we undertake a comparison of the alternative frameworks by estimating ordered and unordered response models using data generated through ordered, unordered data and a combination of ordered and unordered data generation processes. We also examine the influence of aggregate sample shares on the appropriateness of the modeling framework. Rather than be limited by the aggregate sample shares in an empirical dataset, simulation allows us to explore the influence of a broad spectrum of sample shares on the performance of ordered and unordered frameworks. We also extend the data generation process based analysis to under reported data and compare the performance of the ordered and unordered response frameworks. Finally, based on these simulation exercises, we provide a discussion of the merits of the different approaches. The results clearly highlight the emergence of the generalized ordered logit model as a true equivalent ordered response model to the multinomial logit model for ordinal discrete variables.

1. Introduction

Discrete choice models and their variants are employed extensively for analyzing decision processes in various fields including transportation, marketing, social sciences, bio-statistics and epidemiology. Discrete choice models in their broadest sense can be characterized as ordered and unordered response frameworks. The ordered response frameworks are suited for examining discrete variables that are ordinal in nature while the unordered response frameworks are applicable to analyzing all discrete variables. The ordered response models represent the decision process under consideration using a single latent propensity. The choice probabilities are determined by partitioning the uni-dimensional propensity into as many categories as the dependent variable alternatives through a set of thresholds. Examples of ordered discrete variables in the field of transportation include: (1) driver and passenger injury severity in traffic collisions, (2) household vehicle (automobile and bicycle) ownership, and (3) activity participation indicators (such as number of tours, number of stops, activity episode participation frequency and activity duration). The prevalent mechanism to analyze ordered discrete variables is to employ the ordered response models such as ordered logit and ordered probit models depending on the distributional assumptions of the unobserved component of the latent propensity.

Unordered discrete choice frameworks offer a potential alternative to the analysis of ordered discrete variables. These models are characterized, usually, by a latent variable per alternative and an associated decision rule. The most commonly employed unordered discrete models – the multinomial logit (MNL) model and its extensions – have their origin in the random utility domain. The latent variable per alternative is referred to as the alternative utility and the alternative with the highest utility is designated as the chosen alternative. There are a number of studies that have considered the multinomial logit (and its extensions) for examining ordinal discrete choice variables. For example, (1) injury severity (see Yasmin and Eluru, 2012; Savolainen et al., 2011; Eluru and Bhat, 2007 for detailed literature reviews on severity models), and (2) vehicle ownership (see Anowar et al., 2012 for a list of studies).

The applicability of the two frameworks for analyzing ordinal discrete variables has evoked considerable debate on using the appropriate choice model for analysis. There has been considerable debate more recently in the safety community in adopting the appropriate framework for analysis in the injury severity context. There are many strengths and weaknesses for the ordered framework vis-à-vis the unordered framework. The ordered response models explicitly recognize the inherent ordering within the

* Corresponding author. Tel.: +1 514 398 6823; fax: +1 514 398 7361.
E-mail address: naveen.eluru@mcgill.ca

0001-4575/$ – see front matter
http://dx.doi.org/10.1016/j.aap.2013.02.012

decision variable whereas the unordered response models neglect the ordering or require artificial constructs to consider the ordering (for example the ordered generalized extreme value model (OGEV)). On the other hand, the traditional ordered response models restrict the impact of exogenous variables on the choice process to be same across all alternatives while the unordered response models allow the model parameters to vary across alternatives (see Eluru et al., 2008 for a discussion). The restricted number of parameters ensures that ordered response models have a parsimonious specification. The unordered response models might not be as parsimonious but offer greater explanatory power because of the additional exogenous effects that can be explored. In fact, several studies highlight the advantages of multinomial logit models over the ordered response models (see for example Bhat and Pulugurta, 1998).

Another concern with the ordered response framework is in the context of modeling datasets that might be affected by under reporting[1] – an aspect of great relevance to safety literature. In fact, unordered response framework is considered to be more effective compared to the ordered response framework in this context. In the case of an under reported decision variable, the traditional multinomial logit model provides estimates that are unbiased i.e. the elasticity effects of the variables are not affected by the under reported data. This is quite critical in terms of examining exogenous variable impacts on the decision process. Further, the unordered response model can be applied by altering the constants if the true population shares are available. In the case of an ordered response model, the parameter estimates are expected to be biased and hence might lead to erroneous policy implications[2]. In summary, in the context of accident literature there are two important aspects that need to be examined.

(1) Which model framework offers superior statistical fit (and thereby behavioral interpretability) to the dataset under consideration?
(2) How do these frameworks perform in the presence of under reported data?

It is in this light that we undertake the current research effort. The objective of the current study is to investigate the performance of the ordered and unordered response frameworks at a fundamental level. Towards this end, we undertake a comparison of the alternative frameworks by estimating ordered and unordered response models using data generated through ordered, unordered data and a combination of ordered and unordered data generation processes (more on this in Section 2.1.1). Subsequently, we examine the influence of aggregate sample shares on the appropriateness of the modeling framework i.e. are ordered response frameworks more suitable to examine ordered discrete variable with a particular share. Rather than be limited by the aggregate sample shares in an empirical dataset[3], simulation allows us to explore the influence of a broad spectrum of sample shares on the performance of ordered and unordered frameworks. Third, we extend the data generation process based analysis to under reported data and compare the performance of the ordered and unordered response frameworks (in the context of different data generation processes and varying sample shares). Finally, based on these simulation exercises, we provide a summary of the strengths and weaknesses of the two frameworks for analyzing ordered discrete variables in general and for injury severity modeling in particular.

The remainder of the paper is organized as follows. Section 2 provides a background for the proposed research methodology; highlights the motivation for our research and discusses the experimental setup of our study. Section 3 briefly outlines the econometric frameworks of the three alternative frameworks considered. Section 4 presents the results from the simulation exercise for model comparisons. In Section 5, we discuss the simulation results in the context of datasets with under reporting. Section 6 provides a discussion of findings from our analysis while simultaneously providing guidelines on the appropriateness of the modeling frameworks for ordinal discrete variables. We conclude the paper in Section 7 with a discussion of the limitations of the current study and directions for future research.

2. Background and current study in context

2.1. Earlier research

To be sure, some of the aspects highlighted above have been examined in earlier research. For example, Bhat and Pulugurta (1998) undertook a comparison exercise of vehicle ownership decisions through the ordered logit and the multinomial logit models. In their study, the authors estimated the two models on four datasets and confirmed that the multinomial logit model offers superior data fit and validation capabilities. The study highlight how an ordered response model offers a parsimonious specification while the unordered response model offers enhanced behavioral interpretability through the addition of exogenous variable effects at the alternative level. Yamamoto et al. (2008) conducted an analysis of potential under reported data by comparing the performance of ordered probit and sequential binary probit models. The authors found that the sequential probit models outperform the ordered probit model in terms of bias values in the parameters. Ye and Lord (2011) compared the ordered probit, multinomial logit and mixed logit model in terms of under reported data. The authors concluded that all the three models considered in the study perform poorly in the presence of under reported data. The exact impact of under reporting on these model frameworks needs further investigation. The study employed data simulation; however, the models were estimated with just one parameter and for a particular aggregate sample share.

More recently Patil et al. (2012) demonstrated the application of a conditional maximum likelihood estimation approach to address under reporting in the context of crash severity analysis using nested logit model. Anowar et al. (2012) undertook a comparison of the ordered and unordered response models in the context of vehicle ownership. In their study, they found that the unordered response models outperform the ordered response models. Yasmin and Eluru (2012) undertook a comprehensive comparison of various modeling frameworks including ordered logit, generalized ordered logit, multinomial logit, nested logit and ordered generalized extreme value logit for analyzing driver injury severity. In their study, the results clearly establish the superiority of the generalized ordered logit in the context of driver injury severity. In the study the authors also explored the issue of how different frameworks perform in the presence of under reporting in the data. The authors computed elasticity effects for the "true" and under reported datasets and concluded that the error in elasticity effects estimated from the unordered systems is not any better than the error in elasticity effects estimated from the ordered systems in their empirical context.

[1] Injury severity reporting is considered to be substantially affected by under reporting (see Elvik and Mysen, 1999; Yamamoto et al., 2008).

[2] Of course, the *true* advantage of the multinomial logit model in the context of under reporting is slightly reduced because the availability of "true" population level measures to the analyst is quite often rare.

[3] An empirical dataset provides a single realization of the aggregate sample share; thus limiting us in exploring the performance difference of the two frameworks as a function of aggregate sample share.