Journal of Choice Modelling xxx (2017) 1–18

Contents lists available at ScienceDirect

Journal of Choice Modelling

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

A model for broad choice data

David Brownstone [a,*], Phillip Li [b]

[a] Department of Economics, University of California, Irvine, 3151 Social Science Plaza, Irvine CA 92697-5100, United States
[b] Office of Financial Research, US Department of the Treasury, 717 14th St NW, Washington, DC 20005, United States

ARTICLE INFO

Keywords:
Alternative aggregation
Constrained maximum likelihood
Bayesian methods

ABSTRACT

This paper analyzes a discrete choice model where the observed outcome is not the exact alternative chosen by a decision maker but rather the broad group of alternatives which contain the chosen alternative. The model is designed for situations where the choice behavior at a particular level is of interest but only broader level data are available. For example, consider analyzing a household's choice for a vehicle at the make-model-trim level but only choice data at the make-model level are observed. The proposed model is a generalization of the multinomial logit model and collapses to it when there is full observability of the exact choices. We show that the parameters in the model are at least locally identified, but for certain configurations of the data, they are only weakly identified. Methods to address weak identification are proposed when there are data available on the overall market shares of all alternatives, and both maximum likelihood and Bayesian estimation methods are explored.

1. Introduction

Discrete choice models are usually estimated with data on the exact choices made by the decision makers from a well-specified choice set, as well as with observable attributes that are related to the choices, decision makers, or both. With these data standard discrete choice models like multinomial logit, probit, and generalized extreme value models can easily be estimated.

In contrast to this standard setting, our paper focuses on the situation where the econometrician does not observe the choices made by the decision makers at the level of interest, but rather only observes the broad groups of alternatives in which the chosen alternatives belong to. We refer to the choices at the original level of interest as exact choice data and the broader level group data as broad choice data. As a running example, suppose that it is of interest to model a household's vehicle choice at the make-model-trim level where the choice set contains a Honda Civic LX, Honda Civic Hybrid, Toyota Camry LE, and Toyota Camry XLE Hybrid. Instead of observing the household's exact choices from this four-vehicle choice set, the econometrician only observes the broad make-model group choices from the choice set, either from the Honda Civic or Toyota Camry group. The main objective of our paper is to only use the broad choice data to make inferences for the parameters belonging to the original exact choice data (e.g., alternative-specific constants).

There have been a few directions in the literature to address this data observability issue. From the statistics side, this type of data is referred to as either grouped data (Heitjan, 1989; Gjeddebaek, 1956a,b, 1961, 1949, 1957, 1959), partially categorized data (Blumenthal, 1968; Nordheim, 1984), or coarse data (Heitjan and Rubin, 1990, 1991). These three concepts are closely related and generally address the problem of only observing power sets from the sample space for the original random variable of interest. They differ only in terms of the type of random variable being analyzed: grouped data generally refer to observing interval data from continuous random variables, partially categorized data refer to observing set data from discrete random variables, and coarse data refer to observing

* Corresponding author.
E-mail addresses: dbrownst@uci.edu (D. Brownstone), phillip.li@ofr.treasury.gov (P. Li).

https://doi.org/10.1016/j.jocm.2017.09.001
Received 22 May 2017; Received in revised form 15 August 2017; Accepted 13 September 2017
Available online xxxx
1755-5345/

general power set data from any random variable. Our definition of broad choice data is similar to partially categorized data and is similar to coarse data for discrete random variables.

Another direction of research is to redefine both the choice and attribute data into a common level of observability so that standard methods can be applied. For instance, in the vehicle choice example, the observable attributes at the make-model-trim level are either aggregated or averaged into the make-model level prior to estimation (e.g., average Honda Civic miles per gallon is used instead of specific trim-level fuel consumption). The attributes and choice data at the matching make-model level are then analyzed using standard discrete models. This approach has two major drawbacks. One is that using aggregate or average attributes will result in loss of precision for the parameter estimates when the members within a make-model level group are not homogeneous with respect to their attributes. This is obvious since the miles per gallon ratings are significantly different between Honda Civic hybrids and non-hybrids, so averaging over this attribute within the make-model set will create measurement error which will lead to inconsistent parameter estimates. McFadden (1978) shows that if the distribution of attribute values being aggregated can be approximated by a multivariate normal distribution, then this inconsistency can be removed by including the covariances of the attributes within the group as well as the log of the number of alternatives in the group as additional explanatory variables. The second drawback is that, by averaging over the make-model-trim level attributes, there may not be enough variation to identify the parameters specific to the make-model-trim level, which is the level that we wish to make inferences in. We may need to identify these parameters to analyze the impacts of fuel economy standards.

Multiple imputation is another direction of research. Intuitively, this approach imputes the exact choices from the original choice set of interest for each decision maker, estimates the model using the imputed exact choice data and attributes, and averages the parameter estimates over the numerous sets of imputed data. This is an attractive method since, given each set of imputed exact choices, standard discrete choice models can be used. Unfortunately a key requirement for multiple imputation estimators to be consistent is that the estimator must be consistent for each completed data set based on a single set of imputations (Rubin, 2004, Chapter 4). Unless the imputed alternative is the one actually chosen by the household, then the estimates on each completed set of data are not consistent.

In this paper, we propose a formal regression-based model for broad choice data that addresses the drawbacks in the current literature. In particular, our model is different than the work from the statistics literature in that it is a discrete choice model (i.e., the probabilities are based on utility maximization) and is a regression-based (i.e., attribute or covariate-based) model, while most of the current literature in statistics is based on general analysis of the data observability mechanism without covariates. In our framework, the broad choice data can be used together with the attributes at the exact choice level, avoiding the need of the previous literature to redefine data into a common level prior to estimation. The estimators we propose are either maximum likelihood or Bayes estimators, so they are fully efficient.[1] And finally, our paper is unique in that we closely analyze the issue of identification when broad choice data are used instead of exact choice data. We show that the parameters in our model are locally identified, but for certain cases the parameters are only weakly identified (i.e., the likelihood function is almost completely flat). To address this weak identification issue, we introduce a novel technique to incorporate external information into the model in the form of parameter constraints or informative priors (in the Bayesian sense), and we also show how this information can be easily incorporated into maximum likelihood and Bayesian estimation routines. We only consider the case where the underlying choice model is conditional logit, but extensions to other discrete choice models are straightforward. Wong et al. (2017) provide Monte Carlo results showing that the broad choice model described in this paper performs much better than McFadden's procedure, averaging over aggregated alternatives, or using a "representative alternative" in a realistic vehicle choice situation.

The paper proceeds as follows. The model for broad choice data is formally stated in Section 2, and the likelihood-based quantities are derived in Section 3. Using the quantities from the preceding section, Section 4 discusses the identification issues associated with using the broad choice data. The details for maximum likelihood and Bayesian estimation of the parameters are discussed in Section 5, and Section 6 illustrates the various estimators on simulated data. Concluding remarks are in Section 7.

2. Model for broad choice data

The model specification is similar to that of a multinomial logit model and is based on random utility theory. Formally, the model is expressed as

$$U_{ij}^{*} = \delta_j + x_{ij}'\beta + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} \text{Type 1 Extreme Value}, \tag{1}$$

$$Y_i^{*} = j \quad \text{if} \quad U_{ij}^{*} \geq U_{ik}^{*} \; \forall \, k \in C = \{1, 2, \ldots, J\}, \tag{2}$$

$$Y_i = m \quad \text{if} \quad Y_i^{*} \in C_m, \tag{3}$$

for decision makers $i = 1, \ldots, N$, alternatives $j = 1, \ldots, J$, and groups $m = 1, 2, \ldots, M$.

The latent utility that decision maker i obtains from alternative j is given by U_{ij}^{*} in (1). It is a function of an "average" level of utility that is constant for alternative j across all decision makers, δ_j, a column vector of K exogenous and observable attributes, x_{ij}, a column vector of unknown coefficients, β, and an unobserved error term, ε_{ij}, that is distributed i.i.d.Type 1 Extreme Value. For identification

[1] Other estimators such as maximum entropy or M-estimators may also be efficient, but these alternatives are rarely used in applied work.