Journal of Choice Modelling 26 (2018) 28–40

Contents lists available at ScienceDirect

Journal of Choice Modelling

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

Computational methods for estimating multinomial, nested, and cross-nested logit models that account for semi-aggregate data

Jeffrey P. Newman [a], Virginie Lurkin [b], Laurie A. Garrow [c,*]

[a] Cambridge Systematics, Inc., 115 South LaSalle Street, Suite 2200, Chicago, IL, 60603, USA
[b] École Polytechnique Fédérale de Lausanne, Route Cantonale, 1015, Lausanne, Switzerland
[c] Georgia Institute of Technology, 790 Atlantic Drive, Atlanta, GA, 30332-0335, USA

ARTICLE INFO

Keywords:
Discrete choice models
Semi-aggregate data
Airline itinerary choice models

ABSTRACT

We present a summary of important computational issues and opportunities that arise from the use of semi-aggregate data (where the explanatory data for choice scenarios are not necessarily unique for each decision-maker) in discrete choice models. These data are encountered with large transactional databases that have limited consumer information, a common feature in some transportation planning applications, such as airline itinerary choice modeling. We developed a freeware software package called Larch, written in Python and C++, to take advantage of these kind of data to greatly speed the estimation of discrete choice model parameters. Benchmarking experiments against Stata (a commonly used commercial package), Biogeme (a commonly used freeware package), and ALOGIT (a highly specialized commercial package for discrete choice modeling) based on an industry dataset for airline itinerary choice modeling applications shows that the size of the input estimation files are 50–100 times larger in Stata and Biogeme, respectively. Estimation times are also much faster in ALOGIT and Larch; *e.g.*, for a small itinerary choice problem, a multinomial logit model estimated in ALOGIT or Larch converged in less than one second whereas the same model took almost 15 seconds in Stata and more than three minutes in Biogeme.

1. Introduction

Discrete choice models are the backbone of empirical analysis in many fields, including transportation, economics, marketing, public policy and operations research. As its name suggests, "discrete choice models" are used to model how individuals select one (or in some cases more than one) discrete alternative from a set of mutually exclusive and collectively exhaustive alternatives (Ben-Akiva and Lerman, 1985; Train, 2003). An early mathematical form of a discrete choice model was introduced by Daniel McFadden, who in 1972 used a multinomial logit (MNL) model to forecast ridership for the Bay Area Rapid Transit (BART) system (McFadden, 2001). The MNL model is still widely used in various applications because of its mathematical elegance and simplicity, although it is often criticized for employing unrealistic assumptions about behavior. Over the years since the development of the MNL, researchers have derived and estimated dozens of other discrete choice models that have relaxed one or more restrictions associated with the MNL model. For example, the nested logit (NL) (McFadden, 1978; Williams, 1977) and cross-nested logit (CNL) (Vovsha, 1997) incorporate more realistic substitution patterns by relaxing the assumption that error terms are independent across

* Corresponding author.
E-mail addresses: jnewman@camsys.com (J.P. Newman), virginie.lurkin@epfl.ch (V. Lurkin), laurie.garrow@ce.gatech.edu (L.A. Garrow).

https://doi.org/10.1016/j.jocm.2017.11.001
Received 1 February 2017; Received in revised form 21 September 2017; Accepted 1 November 2017

alternatives. The probit (Daganzo, 1979) and mixed logit models (Train, 2003) incorporate random taste variation and can be used for applications in which error terms are correlated across observations.

The objective of this paper is to describe some important computational methods that are especially useful for estimating parameters for choice models with semi-aggregate data (in which the choice scenarios are not necessarily unique across decision-makers). This data feature is commonly encountered with large transactional databases that have limited consumer information. For example, many online retailers have information about the products offered for sale at a given point in time and which of these products were purchased, but do not have information about the consumer who purchased a product. This results in semi-aggregate data in the sense we can aggregate choices for a particular choice set (since we do not have customer-level information). The methods we will describe have been implemented in Larch, a free open-source software package written in Python and C++ that estimates MNL, NL, and CNL models. Larch helps address an emerging problem that many researchers working with large transactional databases face by leveraging more efficient data storage and computational procedures to reduce estimation times for large datasets.

The remainder of this paper describes in detail a number of factors that are relevant for the estimation of discrete choice models for semi-aggregate data. The next section provides an overview of the data (and modeling problem) that motivated the development of Larch. Then we present an overview of some of the discrete choice models that can be estimated efficiently using Larch. The next sections review common data formats (including the identification of chosen and available alternatives), present one method that can be used to number alternatives for large datasets when estimating discrete choice models that contain nests, and describe the methodology used by Larch to weight the log-likelihood function. Finally, we present results from computational experiments based on an industry dataset for airline itinerary choice modeling applications.

2. Problem motivation and data

Itinerary choice models are used to predict the probability that a passenger purchases a specific airline itinerary. Itinerary choice models are used by airlines, aircraft manufacturers, government agencies, etc. and support long-term decisions including where and when to schedule flight legs, how many aircraft to purchase (or manufacture), and which airlines are potentially good code share partners. In itinerary choice models, it is common to construct the set of available itineraries from leg schedule files and/or from revealed purchase transaction data. In the first case, different rules are used to determine which legs can be combined to form multi-leg itineraries. For example, these rules determine minimum and maximum connection times and determine whether legs can be operated by different carriers. In the second case, actual ticket purchases over a longer period (typically a month) are used to construct the set of available alternatives. Although this is not ideal from a theoretical perspective (as itineraries that are infrequently chosen may not be included, potentially leading to bias in parameter estimates), the sheer volume of the ticket transactions helps mitigate this concern (that is, the larger the transaction database, the more likely infrequently chosen itineraries will appear).

The data used for this study were derived from a ticketing database provided by the Airlines Reporting Corporation. The data represent ten origin destination pairs for travel in U.S. continental markets in May of 2013. Itinerary characteristics have been masked, *e.g.*, carriers are labeled generically as "carrier X" and departure times have been aggregated into categories (morning, afternoon, evening). An average fare is provided but is not accurate (a random error has been added to each fare). These modifications were made to satisfy nondisclosure agreements, so that the data used in this paper can be published for teaching and demonstration purposes. It is generally representative of real itinerary choice data used in practice, and the results obtained from this data are intuitive from a behavioral perspective, but it is not accurate and should not be used for behavioral studies.

There are three characteristics of itinerary choice models that heavily influenced our decision to develop Larch. These characteristics include: (1) the number of alternatives in a choice set, (2) the inclusion of different markets in the estimation dataset, and (3) the semi-aggregate nature of the data, in which the choice scenarios are not necessarily unique across decision-makers. With regard to the first point, the number of alternatives in itinerary choice models is large, *e.g.*, Coldren (2005) reports hundreds of itineraries for the estimation dataset used by United Airlines. The maximum number of alternatives we have in our representative dataset is similarly large, as we include as many as 127 alternatives. Although this may not be an especially large choice set compared to certain other contexts (*e.g.*, in residential location choice, there may be thousands of individual alternatives modeled), when considered jointly with the number of observations found in transactional databases (millions, or more), even a few hundred alternatives can be computationally challenging to process. Second, with regards to market segmentation, it is common in itinerary choice models to define a segment as all origin-destination (OD) pairs that share one or more common characteristics. For example, in our earlier work (Lurkin et al., 2017) we defined market segments as a function of distance, direction of travel, and the number of time zones traveled. The inclusion of multiple OD pairs in the same estimation dataset causes some challenges on how to identify nests that share the same characteristics. For example, "alternative 10" in the first OD pair may be operated by American Airlines whereas "alternative 10" in the second OD pair may be operated by United Airlines. Carefully numbering alternatives (described in Section 5) allows us to associate the same set of alternative numbers with itinerary characteristics, but further explodes the number of virtual alternatives in the nest. Finally, because we have no socio-demographic information and because the set of available choices does not vary across individuals, it is very common in our database to have a situation where many individuals face a completely identical choice set, and where some individuals chose the first alternative, some the second alternative, etc. Many existing software applications require that the choice set be "repeated" for each chosen alternative, *e.g.*, that one choice set be defined for those individuals selecting the first alternative, a second (and identical) choice set be defined that for those individuals selecting the second alternative, etc.