Transportation Research Part B 58 (2013) 170–184

Contents lists available at ScienceDirect

Transportation Research Part B

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

Estimating GEV models with censored data ☆

Jeffrey P. Newman [a,*], Mark E. Ferguson [b], Laurie A. Garrow [a]

[a] Department of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, United States
[b] Moore School of Business, University of South Carolina, Columbia, SC, United States

ARTICLE INFO

Keywords:
Discrete choice
GEV
Censored data

ABSTRACT

We examine the problem of estimating parameters for Generalized Extreme Value (GEV) models when one or more alternatives are censored in the sample data, i.e., all decision makers who choose these censored alternatives are excluded from the sample; however, information about the censored alternatives is still available. This problem is common in marketing and revenue management applications, and is essentially an extreme form of choice-based sampling. We review estimators typically used with GEV models, describe why many of these estimators cannot be used for these censored samples, and present two approaches that can be used to estimate parameters associated with censored alternatives. We detail necessary conditions for the identification of parameters associated exclusively with the utility of censored alternatives. These conditions are derived for single-level nested logit, multi-level nested logit and cross-nested logit models. One of the more surprising results shows that alternative specific constants for multiple censored alternatives that belong to the same nest can still be separately identified in nested logit models. Empirical examples based on simulated datasets demonstrate the large-sample consistency of estimators and provide insights into data requirements needed to estimate these models for finite samples.

1. Introduction

Current estimation methods for discrete choice models generally assume that all alternatives are observed to have been chosen for at least some observations in the estimation dataset. The simplest estimators are derived from an assumption that the sample of observations represents a purely random selection of possible observations (i.e., the population). However, using a purely random sample is not always desirable or possible. For example, using a purely random sample of travelers may be undesirable when modeling use of low ridership modes, because a prohibitively large sample would be required to ensure that sufficient quantities of users are sampled to accurately model their preferences. Moreover, in oligopolistic markets, it may even be deemed illegal collusion for competitors to share data about customers. The latter motivates the goal of this paper: to develop estimators for discrete choice models in which one or more alternatives is never observed to have been chosen in the estimation dataset; however, information about the censored alternatives is still available. This problem can be viewed as an extreme case of non-random sampling for the estimation data.

Accommodating stratified samples, where the selection of observations in the sample is not purely random, can be roughly divided into two categories: exogenous samples, where the probability of an observation being sampled is related to some attributes of the alternatives or the decision makers but unrelated to the observed choice; and endogenous samples,

☆ Presented at IATBR Toronto, July 2012; Submitted to Transportation Research Part B, November 2012, revised April and July 2013.
* Corresponding author. Tel.: +1 (312)278 3794.
E-mail addresses: jpn@gatech.edu (J.P. Newman), mark.ferguson@moore.sc.edu (M.E. Ferguson), laurie.garrow@ce.gatech.edu (L.A. Garrow).

0191-2615/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trb.2013.09.002

where the probability of an observation being sampled is related directly to the observed choice. This second case is often called "choice-based" stratified sampling.

A number of modifications to the basic maximum likelihood estimation procedure have been proposed in the literature to accommodate choice-based stratified samples, both for situations where the market shares for the various alternatives are known, and for situations where they are not known. When shares are known, Manski and Lerman (1977) showed it is possible to employ weighted exogenous sample maximum likelihood (WESML), which provides consistent estimators. However, WESML can incur a substantial loss in estimator efficiency (i.e., the standard errors of the estimates are large), notably when the variance in the weights on the observations is large. It is also possible to estimate parameters with choice-based sampling by using conditional maximum likelihood (CML), proposed by Manski and McFadden (1981). The CML methodology can even be used when market shares are unknown and must be estimated alongside the other model parameters (Hsieh et al., 1985). In the case of a multinomial logit (MNL) model with a full set of alternative specific constants, when the market shares are known the CML method reduces to ESML with post-hoc adjustments to the estimated constants. But when market shares and relative sample rates for the various alternatives are not known, the independence of irrelevant alternatives (IIA) property of the MNL model ensures that while consistent estimators for other parameters are available, the true market shares of the alternatives are unidentifiable.

1.1. Censored data

We examine an extreme form of choice-based sampling: instances where one or more of the alternatives is systematically excluded from the sample used to estimate parameters, i.e. the sampling probability for those alternatives is zero. We term this condition "censored" sampling, the resulting sample as "censored data", and the alternative[s] that have zero sampling frequency as "censored alternatives". Under more typical choice based sampling conditions, the probability of individual decision makers being included in the sample is a function of the observed choice, and while some choices result in a smaller probability of being included than others, all decision makers have a non-zero chance of being sampled, and all possible choices are ultimately represented in the sample. With censored data, this is not the case.

Censored data, as we define it here, does not mean that no information about the censored alternatives is available or collected. It merely means that decision makers who choose a censored alternative are never sampled. Importantly, when a decision maker who selects one of the other (uncensored) alternatives is sampled, it is still possible to observe or construct the attributes of both the chosen and non-chosen alternatives, including the censored alternatives. For example, if automobile users are censored in a mode choice model, that means that no automobile users appear in the sample, but the hypothetical travel times and costs for automobile travel can still be computed for users of other modes of travel. This is not substantially different than would be necessary for those observations even if auto users were not censored.

This type of censored data can arise in a variety of contexts. In the transportation planning context, censored data might arise from data collection constraints, such as limited funding or an oversight in survey design. For example, a travel survey might have been conducted which ignored bicycle users, but during subsequent modeling applications policy makers might suddenly feel that bicycling is related to their policy goals and that it should be included in the models. Censoring is particularly common in revenue management contexts, where a firm in a competitive marketplace is attempting to set price and availability of products so as to maximize profit. In that case, the data can be censored because observations of purchase decisions of the firm's own products are readily available, but observations of purchases of competitors' products are not, and for competitive or legal reasons that information may never be available. Moreover, some potential customers may choose to not purchase any product at all this choice is referred to as the "no purchase" alternative, or the "outside good".

Recent works in revenue management (Talluri and van Ryzin, 2004; Vulcano et al., 2010, 2012) and transportation planning (Newman et al., 2012, 2013) have examined the censored data estimation problem, and proposed methodologies to estimate discrete choice model parameters, including alternative specific constants and other alternative specific parameters for censored alternatives. Most of the work on parameter estimation with censored data has been focused on the MNL model because of the convenient mathematical properties of this model. However, it has been shown that if the estimated choice model is MNL, the IIA property prevents the identification of alternative specific constants (or other alternative specific parameters) for censored alternatives, unless some external information (beyond the sample of choice observations) is available (Newman et al., 2012). The outside information can be (but does not necessarily need to be) known market shares for the observable and censored alternatives. It could also be an assumption of a constant arrival rate of potential customers (Talluri and van Ryzin, 2004), the known market share of just the censored alternatives (Vulcano et al., 2012) or an unknown total market size that is assumed to be stable over time (Newman et al., 2012). For certain other choice models, no outside data is required, as has been demonstrated for the nested logit (NL) model (Newman et al., 2013). No outside data is required for the more general models we consider in this paper, as well. Intuitively, this is because the inclusion of covariance terms results in a system of equations that allows identification of alternative-specific parameters for censored alternatives for particular nesting structures.

Much of the literature on parameter estimation with censored data has focused on the unique nature of the problem. But because censored data is a type of choice-based sampling, it is possible to adapt some existing choice-based sampling parameter estimation techniques to censored data. Nevertheless, care must be taken in selecting appropriate tools. As we will outline in Section 3, not all choice-based sampling methodologies will work with censored data.