Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 23 (2017) 758–771

Transportation Research Procedia

www.elsevier.com/locate/procedia

22nd International Symposium on Transportation and Traffic Theory

The initial condition problem with complete history dependency in learning models for travel choices

C. Angelo GUEVARA [a], Yue TANG [b], Song GAO [b,*]

[a]Universidad de Chile, Department of Civil Engineering, University of Chile, Blanco Encalada 2002, Santiago

[b]Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, 141 Marston Hall, 130 Natural Resources Road, Amherst, MA 01003, USA

Abstract

Learning-based models that capture travelers' day-to-day learning processes in repeated travel choices could benefit from ubiquitous sensors such as smartphones, which provide individual-level longitudinal data to help validate and improve such models. However, the common problem of missing initial observations in longitudinal data collection can lead to inconsistent estimates of perceived value of attributes in question, and thus inconsistent parameter estimates. In this paper, the stated problem is addressed by treating the missing observations as latent variables. The proposed method is implemented in practice as maximum simulated likelihood (MSL) correction with two sampling methods in an instance-based learning model for travel choice, and the finite sample bias and efficiency of the estimators are investigated. Monte Carlo experimentation based on synthetic data shows that both the MSL with random sampling (MSLrs) and MSL with importance sampling (MSLis) are effective in correcting for the endogeneity problem in that the percent error and empirical coverage of the estimators are greatly improved after correction. Compared to the MSLrs method, the MSLis method is superior in both effectiveness and computational efficiency. Furthermore, MSLis passes a formal statistical test for the recovery of the population values up to a scale with a large number of missing observations, while MSLrs systematically fails due to the curse of dimensionality. The impacts of sampling size in MSLrs and number of high probability choice sequences in MSLis on the methods' performances are investigated. The methods are applied to an experimental route-choice dataset to demonstrate their empirical application. Hausman-McFadden tests show that the estimators after correction are statistically equal to the estimators of the full dataset without missing observations, confirming that the proposed methods are practical and effective for addressing the stated problem.

Keywords: Endogeneity; initial condition problem; learning model; maximum simulated likelihood; multiple imputaton

[*] Corresponding author. Tel.: +1-413-545-2688 ; fax: +1-413-545-9569.
E-mail address: sgao@umass.edu

10.1016/j.trpro.2017.05.042

1. Introduction

Learning-based models for travel choice capture travelers' learning process in repeated choices (e.g., Ben-Elia and Shiftan; 2010; Lu et al.; 2014; Tang and Gao; conditionally accepted). In a learning model, a traveler' s perception of an alternative's attribute (e.g., travel time) evolves over time based on all her past experience with the alternative. When forming the perception, each past experience with the alternative takes a weight in memory and the perception is a weighted average of all past experience. The weighting scheme of past experience is specific to the learning model in use. Compared to non-learning models where the perception of an alternative is static over time, estimation of a learning model requires data of travelers' complete past experience with the alternatives. Longitudinal data collection in real life, however, inevitably starts midstream, and rarely includes subjects' complete choice histories. Specialized data collection targeted at newcomers (e.g., new employees or students) to a region might provide the needed data, but such efforts are difficult to implement. In the case of incomplete data, the missing initial observations can lead to biased estimate of the perceived value of the attribute in question, and thus inconsistent parameter estimates. Note that the majority of empirical studies on learning models for travel choice are based on experimental data in a laboratory, where subjects make choices from "day" and thus the stated problem does not exist.

An econometric model is said to suffer from endogeneity when the systematic part of the utility is correlated with the error term. The variables that cause the correlation are called the endogenous variables. Endogeneity can lead to inconsistent estimation of model parameters, since changes in the error term are misinterpreted as changes of the endogenous variable. Endogeneity is common in discrete choice models (e.g., probit, logit, nested logit) as the assumption that the explanatory variables are independent from the error term is often violated. Guevara (2010) classifies endogeneity into three types based on their causes: (1) Omission of the variables that are correlated with some observed variables; (2) Simultaneous determination of multiple variables; and (3) The propagation of measurement errors in explanatory variables to the error term. Several correction methods have been developed to solve endogeneity problems (e.g., Guevara and Polanco; 2016; Heckman; 1978; Berry et al.; 1995; Schenker and Welsh; 1988; Brownstone; 1991; Guevara; 2010; Antolin et al.; 2016) . The endogeneity problem this paper tackles can be classified within the third group, a special case in which endogeneity arises because the researcher has an incorrect measure of the attributes of the alternatives perceived by the decision makers.

Solving the initial observation problem for dynamic panel data discrete choice models is known to be a difficult task. Most existing studies deal with first-order Markov process where the dependent variable is only lagged once. The major focus of these studies is that the initial condition is not exogenous due to correlation of error terms over time. Therefore, if there is no serial correlation, first-order Markov process model would not suffer from the problem. For example, Heckman (1981) and Lee (1997) examined the problem of initial conditions in a time-discrete data stochastic process when serially correlated unobservable variables generate the process. Correction methods were proposed and tested with Monte Carlo experiments. More of such studies can be found in the reference list (e.g. Blundell and Bond; 1998; Wooldridge; 2005; Honore and Kyriazidou; 2000; Carro; 2007). In the learning models for travel choice, a current decision depends on the entire history of past experience, defined as a Polya process in Heckman (1981a). The complete history dependence makes the initial condition problem more challenging than those in the existing studies. The model will suffer from the initial observation problem even without serial correlation. To the best of our knowledge, no solution has been developed to date.

In this paper, the proposed method is based on noting that the likelihood function of this problem can be written as a sequence of integrals over the conditional distribution of the possible choices on the missing days. This multifold integral is then maximized using a variation of the maximum simulated likelihood (MSL), which is described in detail by Train (2009). The MSL numerical estimation method has reached great popularity in the past 15 years, thanks to the significant improvement in computational power. This method has been mainly used for the estimation of Logit Mixture models aimed to account for random coefficients or different error component. The application of the method in this paper is different from the usual ones, although all the conditions for consistency described in Train (2009) are extendable, e.g., the need for having the number of draws growing faster than the square root of the sample size. Despite its popularity, the MSL is not exempt from drawbacks. For example, MSL estimators have a downward bias for a finite number of draws, and they may suffer from empirical identification problems, both in the form of false empirical identification and lack of empirical identification. More importantly for this application, MSL may suffer from the problem known as the curse of dimensionality, which in this case implies that the number of draws required