Transportation Research Part B 106 (2017) 173–192

Contents lists available at ScienceDirect

Transportation Research Part B

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

A practical method to test the validity of the standard Gumbel distribution in logit-based multinomial choice models of travel behavior

Xin Ye [a,*], Venu M. Garikapati [b], Daehyun You [c], Ram M. Pendyala [d]

[a] Tongji University, College of Transportation Engineering, Key Laboratory of Road and Traffic Engineering of Ministry of Education, 4800 Cao'an Road, Shanghai, 201804, China
[b] National Renewable Energy Laboratory, Systems Analysis & Integration Section, 15013 Denver West Parkway, Golden, CO 80401, USA
[c] Maricopa Association of Governments, 302 N. First Avenue, Suite 300, Phoenix, AZ 85003, USA
[d] Arizona State University, School of Sustainable Engineering and the Built Environment, 660 S. College Avenue, Tempe, AZ 85287-3005, USA

ARTICLE INFO

Article history:
Received 12 November 2016
Revised 17 October 2017
Accepted 17 October 2017
Available online 8 November 2017

Keywords:
Travel behavior models
Discrete choice models
Violations of distributional assumptions
Test of validity of distributional assumption
Multinomial logit model
Multiple discrete-continuous extreme value model

ABSTRACT

Most multinomial choice models (e.g., the multinomial logit model) adopted in practice assume an extreme-value Gumbel distribution for the random components (error terms) of utility functions. This distributional assumption offers a closed-form likelihood expression when the utility maximization principle is applied to model choice behaviors. As a result, model coefficients can be easily estimated using the standard maximum likelihood estimation method. However, maximum likelihood estimators are consistent and efficient only if distributional assumptions on the random error terms are valid. It is therefore critical to test the validity of underlying distributional assumptions on the error terms that form the basis of parameter estimation and policy evaluation. In this paper, a practical yet statistically rigorous method is proposed to test the validity of the distributional assumption on the random components of utility functions in both the multinomial logit (MNL) model and multiple discrete-continuous extreme value (MDCEV) model. Based on a semi-nonparametric approach, a closed-form likelihood function that nests the MNL or MDCEV model being tested is derived. The proposed method allows traditional likelihood ratio tests to be used to test violations of the standard Gumbel distribution assumption. Simulation experiments are conducted to demonstrate that the proposed test yields acceptable Type-I and Type-II error probabilities at commonly available sample sizes. The test is then applied to three real-world discrete and discrete-continuous choice models. For all three models, the proposed test rejects the validity of the standard Gumbel distribution in most utility functions, calling for the development of robust choice models that overcome adverse effects of violations of distributional assumptions on the error terms in random utility functions.

* Corresponding author.
E-mail addresses: xye@tongji.edu.cn (X. Ye), venu.garikapati@nrel.gov (V.M. Garikapati), dyou@azmag.gov (D. You), ram.pendyala@asu.edu (R.M. Pendyala).

https://doi.org/10.1016/j.trb.2017.10.009
0191-2615/

1. Introduction

The Gumbel distribution (also called Type-I extreme value distribution) plays a central role in modeling travel behavior, be it in discrete choice models (McFadden, 1974) or in multiple discrete-continuous choice models (e.g., Bhat, 2005, 2008). The attractiveness of this assumption can be largely attributed to the following two reasons. First, the Gumbel distribution is very similar to the normal distribution. In the absence of any specific information about the behavioral phenomenon under investigation, econometric choice models often assume that the random error term (which captures the overall impact of unobserved factors) is normally distributed. Second, when the Gumbel distribution is assumed for random components in utility functions, it is possible to obtain a closed-form expression for the likelihood function using the utility maximization principle. A neat closed-form expression for the likelihood function facilitates consistent and efficient estimation of model coefficients using standard maximum likelihood estimation (MLE) methods.

The Multinomial Logit (MNL) model and Multiple Discrete-Continuous Extreme Value (MDCEV) model, both of which are based on the Gumbel distribution assumption for the random error components, are widely used to predict behavioral choices in a number of fields. Although strides have been made in estimating model formulations that assume a normal distribution for the random error components including, for example, the Multinomial Probit Model (Train, 2009) and the Multiple Discrete-Continuous Probit (MDCP) model (Bhat et al., 2013), the logit-based models continue to be the most common model forms of choice for travel behavior modeling, owing to their ease of estimation and application (e.g. You et al., 2014; Garikapati et al., 2014; Wang et al., 2016). However, the theory of maximum likelihood estimation posits that the consistency and efficiency of maximum likelihood estimators depend on the validity of the distributional assumptions made on the random error components. If the distributional assumption is violated, then the maximum likelihood estimators are neither consistent nor efficient, thus contributing to potentially erroneous forecasts and policy impact assessments.

In prediction mode, the MNL model ensures that predicted market shares match the observed shares in the sample (Ben-Akiva and Lerman, 1985). Therefore, in the case of the MNL model, violations of the distributional assumption will not adversely affect the predicted aggregate market shares. In the MDCEV model, however, such a property does not hold. Jäggi et al., (2013) found that predictions from MDCEV models of vehicle fleet composition and usage are quite sensitive to model specification. As the unobserved but significant factors affecting vehicle fleet composition and usage are absorbed into the random error components, they are bound to influence the nature of the distribution of the random error terms. If the model specification results in a situation where there is a violation of the standard Gumbel distributional assumption on the random error terms of the MDCEV model, it is reasonable to expect substantial inaccuracies in model predictions depending on the severity of the violation.

It is therefore important to test the validity of distributional assumptions on the random error terms when applying the MLE method to estimate model coefficients of either discrete or discrete-continuous travel choice models. The objective of this paper is to propose a practical, and yet strict, statistical method to test whether the error terms in random utility functions of MNL or MDCEV models truly follow the (assumed) Gumbel distribution.

2. Literature review

Econometricians have been questioning the validity of the distributional assumption on random error components of utility functions ever since McFadden (1974) first proposed the multinomial logit model formulation (e.g., Manski, 1975). Concerns about violations of distributional assumptions on the random error terms have motivated the development of semi-parametric and semi-nonparametric choice models. The semi-parametric choice model employs the kernel density method to estimate the distribution of the random errors, and therefore does not rely on any parametric distributional assumptions (e.g., Klein and Spady, 1993; Lee, 1995). The semi-nonparametric (SNP) choice model is based on a polynomial approximation of a probability density function (PDF) that takes a flexible form (Gallant and Nychka, 1987). Because the likelihood function has an explicit analytical expression, the SNP choice modeling method appears to be more widely applied in practice than the semi-parametric approach (e.g., Chen and Randall, 1997; Creel and Loomis, 1997; Crooker and Herriges, 2007).

In this paper, the SNP approach is used to derive a statistical test of the validity of the Gumbel distribution in logit models of discrete choice. It is therefore prudent to first review the SNP binary choice model. Similar to the binary probit model, the SNP binary choice model is also based on a random utility (U) function, which can be expressed as $U = V + \varepsilon$, where "V" is the systematic or deterministic component of the utility function and "ε" is the random component. For the sake of notational brevity, the index 'i' corresponding to an individual observation is omitted in the formulation presented here. If a dummy variable "y" indicates whether an alternative is chosen or not, then $P(y=1) = P(U > 0) = P(V + \varepsilon > 0) = P(\varepsilon > -V)$. The probability density function takes the following form:

$$f(\varepsilon) = \frac{\left(\sum_{m=0}^{K} a_m \varepsilon^m\right)^2 \varphi(\varepsilon)}{\int_{-\infty}^{+\infty} \left(\sum_{m=0}^{K} a_m \varepsilon^m\right)^2 \varphi(\varepsilon) d\varepsilon} \tag{1}$$

In Eq. (1), "K" is the length of the polynomial, "m" is an index increasing from 0 to "K", a_m is a constant coefficient, and $\phi(\varepsilon)$ represents the PDF of the standard normal distribution. The denominator ensures that $\int_{-\infty}^{+\infty} f(\varepsilon) d\varepsilon = 1$. Eq. (1) can be