Transportation Research Part B 78 (2015) 222–237

Contents lists available at ScienceDirect

Transportation Research Part B

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

Accounting for stochastic variables in discrete choice models

Federico Díaz [a], Víctor Cantillo [b], Julian Arellana [b,*], Juan de Dios Ortúzar [c]

[a] Transmetro S.A.S., Calle 45 Carrera 46, Barranquilla, Colombia
[b] Department of Civil and Environmental Engineering, Universidad del Norte, Km 5 Antigua Vía a Puerto Colombia, Barranquilla, Colombia
[c] Department of Transport Engineering and Logistics, Centre for Urban Sustainable Development (CEDEUS), Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Santiago, Chile

ARTICLE INFO

Article history:
Received 17 July 2013
Received in revised form 29 April 2015
Accepted 30 April 2015
Available online 22 May 2015

Keywords:
Stochastic variables
Errors in variables
Discrete choice models
Mixed logit

ABSTRACT

The estimation of discrete choice models requires measuring the attributes describing the alternatives within each individual's choice set. Even though some attributes are intrinsically stochastic (e.g. travel times) or are subject to non-negligible measurement errors (e.g. waiting times), they are usually assumed fixed and deterministic. Indeed, even an accurate measurement can be biased as it might differ from the original (experienced) value perceived by the individual.

Experimental evidence suggests that discrepancies between the values measured by the modeller and experienced by the individuals can lead to incorrect parameter estimates. On the other hand, there is an important trade-off between data quality and collection costs. This paper explores the inclusion of stochastic variables in discrete choice models through an econometric analysis that allows identifying the most suitable specifications. Various model specifications were experimentally tested using synthetic data; comparisons included tests for unbiased parameter estimation and computation of marginal rates of substitution. Model specifications were also tested using a real case databank featuring two travel time measurements, associated with different levels of accuracy.

Results show that in most cases an error components model can effectively deal with stochastic variables. A random coefficients model can only effectively deal with stochastic variables when their randomness is directly proportional to the value of the attribute. Another interesting result is the presence of confounding effects that are very difficult, if not impossible, to isolate when more flexible models are used to capture stochastic variations. Due the presence of confounding effects when estimating flexible models, the estimated parameters should be carefully analysed to avoid misinterpretations. Also, as in previous misspecification tests reported in the literature, the Multinomial Logit model proves to be quite robust for estimating marginal rates of substitution, especially when models are estimated with large samples.

1. Introduction

The estimation of discrete choice models requires data such as socioeconomic characteristics of individuals and attributes of the alternatives within their choice sets. These explanatory variables are usually assumed to be inherently deterministic,

* Corresponding author.
E-mail addresses: fediaz@transmetro.gov.co (F. Díaz), vcantill@uninorte.edu.co (V. Cantillo), jarellana@uninorte.edu.co (J. Arellana), jos@ing.puc.cl (J.de Dios Ortúzar).

http://dx.doi.org/10.1016/j.trb.2015.04.013
0191-2615/

that is, that they would yield the same values if measured repeatedly. The problem is that some variables are actually intrinsically stochastic (e.g. travel times under congested conditions[1]) and thus assuming them to be fixed can be fairly heroic. In fact even an accurate measure can be biased if it is different from the value perceived by the decision maker.

Furthermore, variables which are intrinsically non-stochastic can still be measured inaccurately producing measurement errors. These errors induce a particular kind of randomness from the modeller's point of view. For instance, in strategic planning applications it is common practice to use zone-based network models to obtain level of service attributes, such as travel times, instead of measuring this key attribute at an individual level due to the high data collection costs involved. Also, trips with different levels-of-service are usually temporally and spatially aggregated (e.g. the set of trips between two specific zones at a peak hour) and a single level-of-service value (e.g. an "average" value) is assigned to them, which is evidently different from the true values experienced by the users (Train, 1978). Measurement errors also occur when values are directly provided by the individual in a revealed preference (RP) survey (e.g. waiting time to board a bus, income, or preferred departure time). In the latter case the difference between the reported value and the real one can be significant due to cognitive issues or even policy bias (Daly and Ortúzar, 1990).

When a discrepancy between the "true" value and the value measured by the modeller exists, an estimation bias arises as discussed by Ortúzar and Willumsen (2011, Section 9.2). Let us consider a simple Multinomial Logit (MNL) model with a typical utility function $U = \beta x + \varepsilon$, where β are parameters to be estimated, x are measured attributes and ε is an independent and identically distributed Gumbel error term with mean zero and standard deviation σ_ε. Assume there is a difference between the attribute values as perceived by the modeller (x^*) and the true values (x), such that: $x = x^* + \eta$, where η distributes with mean zero (i.e. no systematic bias exist) and standard deviation σ_η. In this case the utility function is transformed to: $U = \beta\,(x^* + \eta) + \varepsilon$, that is: $U = \beta\, x^* + (\varepsilon + \beta\eta) = \beta x^* + \delta$. The outcome of this is that in the original model, the estimated parameter β' would be:

$$\beta' = \frac{\pi}{\sqrt{6}\cdot\sigma_\varepsilon}\beta \tag{1}$$

whilst in the second model the estimated parameter β'' would be[2]:

$$\beta'' = \frac{\pi}{\sqrt{6}\cdot\sigma_\delta}\beta \tag{2}$$

and the standard deviation of the distribution function of the new error component δ would be:

$$\sigma_\delta = \sqrt{\sigma_\varepsilon^2 + \beta^2\cdot\sigma_\eta^2} \tag{3}$$

Hence $\beta'' < \beta'$ and this estimation bias may affect the model forecasts.

There is also experimental evidence about bias estimation and miscalculation of marginal rates of substitution when measurement errors occur. For instance, Train (1978) explores the use of more accurate data in the estimation of mode choice models concluding that it is sometimes advisable to carry out an additional effort to correct for the measurement bias of some attributes, such as transit transfer time, when analysing transport policies. Ortúzar and Ivelic (1987) showed that using very precise real data, measured at the individual level, resulted in better fit and clearly different subjective values of time when estimating mode choice models in comparison with models estimated with aggregate data. More recently, Bhatta and Larsen (2011) show, using synthetic data, how measurement biases may induce biased parameter estimates on a MNL model, besides miscalculation of marginal rates of substitution. Therefore the use of more accurate (but more expensive) data results in better parameter estimates and this clearly establishes a trade-off between data quality and data collection costs (Daly and Ortúzar, 1990).

In this paper we deal with the problem of working with incorrigibly biased data due to the stochastic nature (inherent or not) of some variables. After a brief review of relevant literature in Section 2, we carry out an econometric analysis to identify appropriate specifications to account for stochastic variables in discrete choice models (Section 3). Then, in Section 4 the performance of some specifications arising from the econometric analysis will be tested and compared in terms of parameter estimate bias, computation of marginal rates of substitution and forecasting ability using both, synthetic and real datasets. Finally, Section 5presents our main conclusions.

2. The problem of errors in variables (EIV)

Much of the effort to specify stochastic variables when estimating econometric models has arisen from the need to solve the EIV problem. In this sense, although there is a vast literature in the case of regression models, research underlying EIV within discrete choice models is scarce, but has shown lately some significant progress. For instance, in the fields of biology and medicine, the EIV problem has been explored in the case of binary models, proposing adaptations of maximum likelihood estimators for specific circumstances (Carroll et al., 1984; Stefansky and Carroll, 1985, 1987, 1990; Kao and Schnell,

[1] Related problems arising from the inherent variability of some level-of-service attributes such as travel time are reliability and risk aversion (Jackson and Jucker, 1982). In this research we will only address the difference between the true value and the values measured by the modeller as a result of this variability.

[2] Let us assume that $\delta = \varepsilon + \beta\eta$ also follow an IID Gumbel distribution, just for illustrative purposes.