Transportation Research Part B 117 (2018) 520–541

Contents lists available at ScienceDirect

Transportation Research Part B

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

Quantifying errors in travel time and cost by latent variables

Juan Manuel Lorenzo Varela[a,*], Maria Börjesson[b], Andrew Daly[c]

[a] KTH Royal Institute of Technology/CTS, Stockholm, Sweden
[b] VTI Swedish National Road and Transport Research Institute/CTS, Stockholm, Sweden
[c] ITS, Leeds, United Kingdom

ARTICLE INFO

Article history:
Received 30 January 2018
Revised 29 August 2018
Accepted 14 September 2018

Keywords:
Hybrid choice models
Latent variables
Error quantification
Measurement error models
RP value of time
Self-reported indicators

ABSTRACT

Travel time and travel cost are key variables for explaining travel behaviour and deriving the value of time. However, a general problem in transport modelling is that these variables are subject to measurement errors in transport network models. In this paper we show how to assess the magnitude of the measurement errors in travel time and travel cost by latent variables, in a large-scale travel demand model. The case study for Stockholm commuters shows that assuming multiplicative measurement errors for travel time and cost result in a better fit than additive ones, and that parameter estimates of the choice model are impacted by some of the key modelling assumptions. Moreover, our results suggest that measurement errors in our dataset are larger for the travel cost than for the travel time, and that measurement errors are larger in self-reported travel time than software-calculated travel time for car-driver and car-passenger, and of similar magnitude for public transport. Among self-reported travel times, car-passenger has the largest errors, followed by car-driver and public transport, and for the software-calculated times, public transport exhibits larger errors than car. These errors, if not corrected, lead to biases in measures derived from the models, such as elasticities and values of travel time.

1. Introduction

Although considerable research has been devoted to measurement errors in the econometric literature, far less attention has been paid to measurement errors in discrete choice modelling and transportation. Recent studies in the transportation field have shown that when measurement errors exist in discrete choice models, the explanatory variable becomes correlated with the error term, and endogeneity problems arise, analogous to those of their linear counterparts. How measurement errors give rise to endogeneity is discussed extensively in Walker et al. (2010); Díaz et al. (2015); Vij and Walker (2016). To reduce bias arising from measurement errors, statistical models that can be used to accommodate errors in explanatory variables have become increasingly popular. These methods include but are not limited to the Control-Function method; the use of Instrumental Variables; the Multiple Indicator Solution; or the integration of Latent-Variables (Guevara, 2015); and among them, the Hybrid Choice Model (HCM) is the modern workhorse in discrete choice analysis.

Parameter bias due to measurement errors in input variables has been highlighted as a substantial problem in the appraisal of policy. For instance, there are reasons to expect that travel cost variables have substantial errors, which attenuate the cost parameters in transport models and lead to under-estimation of the response to pricing measures in appraisal. Moreover, errors in the time and the cost variables are one major reason for collecting Stated Preference (SP) data for value

* Corresponding author.
E-mail address: jmlv@kth.se (J.M.L. Varela).

https://doi.org/10.1016/j.trb.2018.09.010
0191-2615/

of time estimation, leading to other problems, such as: reference dependence and gain-loss asymmetry (De Borger and Fosgerau, 2008; Börjesson and Eliasson, 2014; Börjesson and Fosgerau, 2015; Hess et al., 2017).

The aim of this paper is to explore the capabilities of the HCM framework to quantify the magnitude of the errors in the key explanatory variables in large-scale travel demand modelling. Quantification of the measurement errors in input data will help identifying the least reliable variables, aiding modellers to concentrate efforts where they are most needed. Hence, we expect our findings to be of interest not only to discrete choice modellers, but also to transport planning practitioners.

The first application of the HCM to account for measurement errors can be found in Walker et al. (2010), where a latent variable approach is introduced to deal with error-prone travel times. Using the same methodology, Varotto et al. (2017) investigate how the time parameter changes when accounting for measurement errors. Furthermore, Walker et al. (2010) and Vij and Walker (2016) use Monte Carlo experiments to show that the estimated parameters converge to their "true" value as the model accounts for measurement errors in the input variables. However, these studies do not treat cost variables as latent, or model more than one latent variable at a time; and, to date, no study on large-scale transport models has quantified measurement errors in both time and cost variables; important biases therefore could not be detected. Nor do these studies use self-reported travel times and costs, which are a feature of some travel surveys and which give useful additional information about biases.

In this study, we use a HCM to account for measurement errors in the time and cost variables in a large-scale mode choice model estimated on National Travel Survey (NTS) data. First, we explore the sensitivity of parameter estimates to different modelling assumptions. We show that assumptions regarding the distributions of the latent variables and the measurement error impact the estimated error in the latent attributes, and the parameter estimates of the choice model. Second, we show how goodness-of-fit measurements can be used to rank what measurement error model formulation fits the observed data best. Third, we show how the measurement errors in the time and the cost variables can be compared using a multiplicative measurement error formulation. Fourth, we present the policy implications of these findings, including impact on the model elasticities and Values of Travel Time (VoT).

The rest of the paper is structured as follows: Section 2 presents the modelling framework and the modelling assumptions to be tested. Section 3 provides an overview of the data. Section 4 presents the application of the framework in a case study. Section 5 gives some model properties and Section 6 concludes.

2. Methodology

2.1. Hybrid choice model

We take the equation framework of Walker et al. (2010) as our starting point. These authors treat the "true" value of the explanatory variable suffering from measured errors as a latent variable (X), known only up to a distribution $f_X(X;\theta)$, where θ is a set of estimated parameters, and the measured value of the variable (X) is used as an indicator (I). They define a mode choice model, with the choice probability of alternative i conditional on the set of parameters β as

$$P(i|\beta, X). \tag{1}$$

Since X is unknown, it is necessary to integrate the conditional choice probability over the distribution of X such that

$$P(i|\beta,\theta) = \int P(i|\beta, X) f_X(X;\theta) dX. \tag{2}$$

The measurement equation assumes that the distribution of the indicator (I) conditional on X and a set of estimated parameters λ is

$$I \sim f_M(I|X;\lambda). \tag{3}$$

Putting the pieces together, the likelihood function of the choice model is

$$L(i, I|\beta,\theta,\lambda) = \int P(i|\beta, X) f_M(I|X;\lambda) f_X(X;\theta) dX \tag{4}$$

The unknown parameters (β, θ, λ) can be estimated through maximum likelihood estimation using observed choices, observed characteristics of the alternatives and individuals, and the indicator variables.

The model schematics are shown by Fig. 1, where observed variables, indicators and choices are represented by rectangular boxes, whilst unobserved variables such as utilities and latent variables are represented by ellipses. In addition, structural equations are represented by continuous lines and measurement equations by dashed lines.

In this paper we go beyond the work of Walker et al. (2010) by further developing the specification of the model, including both the latent variable distribution (using different distributions) and the measurement relationship between the measured travel time and the latent variable (better capturing the sources of error). We are not aware of any previous work that uses multiple indicators for travel time variables. We show how parameter estimates of the measurement equations can be used not only to assess the goodness-of-fit of the postulated theories, but also to provide insights into the magnitude of the errors in the explanatory variables.

Vij and Walker (2016) has shown how each HCM has a "reduced form" that will give identical model fit and is simpler and faster to estimate. Nevertheless, in this study we refrain from using reduced-form models because they do not have measurement equations; hence, they are of limited use for the purpose of quantifying measurement errors. Moreover, the