ADDRESSING DISTRIBUTIONAL ASSUMPTIONS IN DISCRETE CHOICE MODELLING OF TRAVEL BEHAVIOUR

Vojtěch Máca - Jan Melichar

Abstract

Travel behaviour is one of the pioneering domains of discrete choice modelling. Logit quickly became the common way of estimation of the probability of choosing one of the transportation alternatives. The intricate issues concerning plurality of consumer preferences – not limited to particularities of mode choice only – paved the way for more elaborated and computationally intensive procedures such as random parameter models that relaxed some strict assumptions inherent to conventional logit estimation. Only recently there have been attempts to address another limitation of the conventional discrete choice models, i.e. assumption on the distribution of random terms.

In this paper we explore the magnitude of these limitations by comparing the common estimation procedures with semi-parametric techniques suggested by Fosgerau (2007). To this end we use the data from a recent travel behaviour study conducted on a sample of Czech population travelling between two major cities – Prague and Brno – by any of the three main land transport means – by car, train or bus. We investigate the consequences of the choice of modelling approach on the value of travel time as well as on the value of travel time variability measured by standard deviation of the travel time.

Key words: discrete choice models, logit, semi-parametric methods, travel behaviour.

JEL Codes: C25, C14, R41.

Introduction

In developed countries (as well as in the Czech Republic) larger transport projects are routinely subjected to cost-benefit analysis. Often, the time savings expressed by the value of travel time savings are the major part of the benefit side in the analysis. Consequently, the choice of value of travel time to be used in the analysis is of a great importance and should be therefore well-founded.

The paper presents first results from an empirical study looking at value of travel time savings and reliability conducted as a part of a research project on quantification of external costs of transport in the Czech Republic. The study scope was restricted to trips between two major Czech cities – Prague and Brno – what provides a good opportunity to look on all three major land transport modes that are regularly used on this route – car, train and bus.

1 Data and methods

1.1 Survey and data

The study presented in this paper refers to one of two choice experiments conducted as a part of a travel behaviour survey looking at both revealed and stated preferences of respondents that have undertaken at least one trip between Prague and Brno (or vice versa) over last 30 days. In total 602 questionnaires were collected using computer assisted web interviewing (CAWI) among residents of Prague and Brno agglomerations between November 2010 and April 2011 with roughly proportional shares of the three modes – car, train and bus as well as business and non-business trips in each segment. The recruitment of respondents was conducted by social research company SC&C via interviewers' network, using advertisements and snowballing.

The questionnaire structure followed a common practice (see e.g. Louviere et al. 2000) starting with the part asking about last trip between the two cities of interest, followed with stated choice experiment pivoted around the last trip characteristics so as to render the scenarios presented realistic. In the last part socio-demographic and other possible explanatory characteristics of the respondent and his/her household were surveyed.

The choice experiment used a travel costs attribute and a popular 5-levels presentation of travel time variability using a mean variance approach originally devised by Black and Towriss (1993) and frequently used in subsequent studies (Small et al. 1999, Ramjerdi et al. 2010).

The choice experiment was designed according to the Bradley design as a choice between two alternatives (denoted Trip A and Trip B) with the advantage of having status quo present in every choice situation. In this experiment respondent faced 9 choice situations in total with 5^{th} choice situation designed as dominant deliberately included to allow for controlling of consistency.

The choice situations were described using 3 attributes – travel costs, travel time and travel time variability but only travel costs and five possible travel times were shown to the respondents. Each of the attributes was assigned one of 5 levels (-2, -1, 0, 1, 2); the levels of travel costs and travel time were set in relation to reference trip described in revealed preference part (i.e. the attributes for level 0 matched exactly the reference trip). Travel time variability levels were pivoted asymmetrically around a base level and setting higher variance for car compared to that for public transport.

1.2 Discrete choice modelling

There are two popular approaches to variability of travel time – mean variance approach and scheduling approach (Small et al. 1999). Our design with a choice between two alternatives described by travel costs and 5 possible travel times is common format for a mean variance approach. The basic model of systematic (indirect) utility for this set-up consists of three variables in linear form, mathematically denoted as

$$V = \beta_M M + \beta_T T + \beta_V \sigma_T \tag{1}$$

where M is travel costs, T is expected travel time and σ_T is standard deviation of travel time, and β's are parameters – marginal utilities of cost, travel time and variability – to be estimated. As the most convenient modelling approach logit estimation (e.g. Ben-Akiva and Lerman, 1985; Train, 2003) is used to determine the probability of choosing one of the two alternatives presented in each of 8 choice sets, i.e.

$$P(i) = \frac{1}{1 + e^{-\mu(V_i - V_j)}} \tag{2}$$

where V_i and V_j denotes utility from choosing alternatives i and j respectively, for linear-in-parameters utilities the parameter μ is assumed to be equal to 1 (Small and Verhoef, 2007) and probability of choosing alternative j is simply equal to 1 – P(i). The most widely used method for estimation of unknown parameters (β's) is maximum likelihood (Ben-Akiva and Lerman, 1985). Defining an indicator variable

$$y_{in} = \begin{cases} 1 \; if \;\; person \; n \; choose \; alternative \; i \\ 0 \; if \;\; person \; n \; choose \; alternative \; j \end{cases} \tag{3}$$

the logarithmic transformation of likelihood function is then written as

$$L(\beta_1,...,\beta_K)=\sum_{n=1}^{N}\left(y_{in}\ln P_n(i)+y_{jn}\ln P_n(j)\right) \tag{4}$$

The most meaningful interpretations of estimated coefficients are their ratios, which express marginal rates of substitution, i.e. between time and money in indirect utility function, what is generally denoted as value of travel time.

In conventional random utility models – multinomial logit in this case – only error term in utility function is introduced to capture randomness in individual behaviour and rest of parameters are deterministic (i.e. value of travel time varies with travel time but otherwise is the same for everyone). However, this assumption is quite restrictive and many individually specific determinants (e.g. socioeconomic, personal traits, features of the travel etc.) are missing in the model. The problem here is that it is not known how the coefficients are distributed in the population, but can in principle be assumed to be random and follow some statistical distribution. Specifically, each individual's coefficient β_n is different from population mean β by some unknown amount, representing taste variance.

Mathematically, the probability that respondent n chooses alternative i can be expressed as

$$P_n(i)=\frac{e^{\beta_n' x_{in}}}{\sum_{j\in C_n} e^{\beta_n' x_{jn}}} \tag{5}$$

where $\beta_n=(\beta_{1n},\beta_{2n},...,\beta_{Kn})'$.

The distribution of absolute values of coefficient is commonly assumed to follow certain known distribution (normal, truncated normal and lognormal being frequently used). The need for imposition of a specific functional form beforehand has been recently addressed in similar studies and applications of alternative techniques were suggested that leave functional form and distributional assumption unspecified (non-parametric) or with only some parametric element (semi-parametric) (Fosgerau, 2007). We use one of these techniques called smoothing splines. The basic model resembles basic ordinary least squares model:

$$y_i = f(x_i)+\varepsilon_i \tag{6}$$

In contrast to OLS, here we are searching for a balance between fit of $f(x_i)$ and smoothness of certain form. A polynomial variant of smoothing splines aims at minimizing the following modified least square criterion (Faraway, 2006):

$$\frac{1}{n}\sum(Y_i - f(x_i))^2 + \int [f''(x)]^2 dx \tag{7}$$

To illustrate the smoothing splines approach we estimate a generalized additive model with logistic link using the R package mgcv (Wood, 2006).

2 Results

All the results were estimated in R statistical software (R-development team, 2011). We report only the results of models with basic variables, i.e. travel time (in minutes), travel costs (in Czech crowns) and standard deviation of 5 possible travel times, from various other predictors tested only business trip was significant in car segment (but at 10% significance level only).

Tab. 1: Multinomial models

segment	**CAR**		**TRAIN**		**BUS**	
predictor	**coefficient**	**std.err.**	**coefficient**	**std.err.**	**coefficient**	**std.err.**
cost	-0.0065***	(0.0005)	-0.0299***	(0.0016)	-0.0312***	(0.002)
time	-0.0208***	(0.0028)	-0.0368***	(0.0032)	-0.0356***	(0.0035)
reliability	-0.0068**	(0.0025)	-0.0128***	(0.0024)	-0.025***	(0.0031)
Log-Likelihood	-1007		-970,77		-915,05	
adj.rho^2	0,137		0,208		0,1466	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '

Source: own calculations

The two values of interest are value of time and reliability ratio given as the estimated coefficient of reliability (i.e. travel time mean variance) divided by travel time (Hollander 2005). The estimated value of time it is the highest for car drivers at CZK 192 per hour, followed by train at CZK 74 per hour and the lowest for bus at CZK 68 per hour. The implicit reliability ratio is also the highest for car drivers (1.04), followed by train passengers (0.8) and lowest for bus passengers (0.43).

Random coefficient models are reported in the table below. Time and reliability coefficients are set as random in the models while cost coefficient is kept fixed, what simplifies the estimation of value of time. Since random coefficient models do not have closed form, simulated maximum likelihood estimator by means of Halton draws is used to estimate the model.

Tab. 2: Random coefficient models

segment	**CAR**		**TRAIN**		**BUS**	
predictor	**coefficient**	**std.err.**	**coefficient**	**std.err.**	**coefficient**	**std.err.**
cost	-0.0109***	(0.0005)	-0.0397***	(0.0025)	-0.0495***	(0.0033)
time	-0.0362***	(0.0042)	-0.0483***	(0.0051)	-0.0583***	(0.006)
reliability	-0.0196***	(0.0035)	-0.0161***	(0.0038)	-0.0429***	(0.005)
sd. time	0.0459***	(0.006)	0.0472***	(0.0071)	0.0512***	(0.008)
sd. reliability	0.0817***	(0.0062)	0.0635***	(0.0061)	0.0876***	(0.0074)
Log-Likelihood	-914,64		-819,52		-798,16	
adj.rho^2	0,2131		0,2452		0,2534	

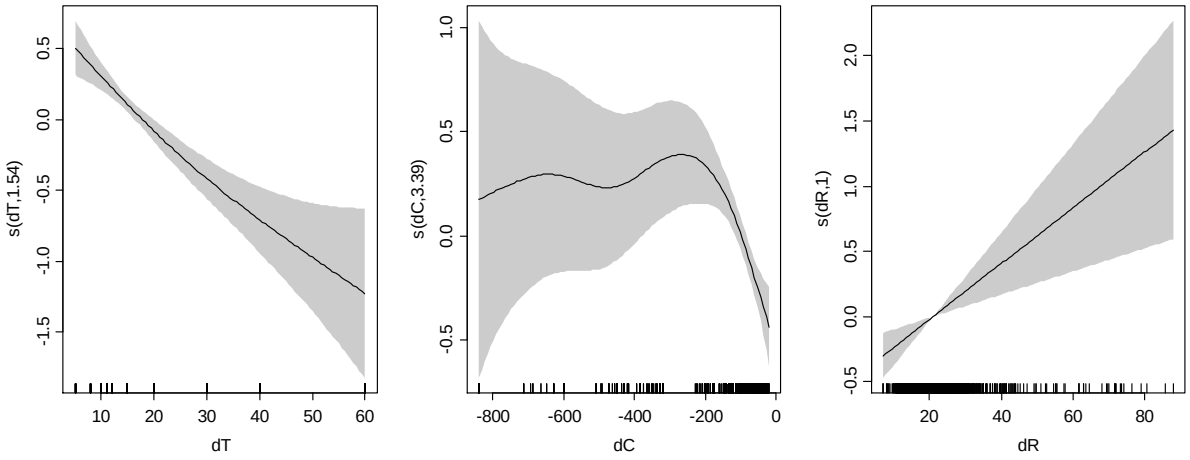
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '

Source: own calculations

The model fit is much better compared to multinomial logit. The implied values of time have however changed only slightly – for car drivers to CZK 199 per hour, for train passengers to CZK 73 per hour and for bus passengers to CZK 71 per hour.

To illustrate the capabilities of semi-parametric model a generalized additive model with logistic link was estimated. The next figure shows the smoothing spline fits with bandwidths for each of the tree predictors – travel time (dT), cost (dC) and reliability (dR), estimated for passenger car segment.

Fig. 1: Semiparametric logit model with additive component functions for cost and time



Source: own calculations

The visualisation of the spline fitted components suggests that the cost coefficient's shape does not resemble any of distributions commonly used in random parameter models;

the deviance explained (all the results available from the authors) is however very low, around 3% only.

Conclusion

The paper explores different models for discrete choice data on intra-mode choices using the data on travel behaviour on the route between Prague and Brno.

Both multinomial and random parameter models are shown to give very similar results. Furthermore, semi-parametric model using splines is employed and visualisation of the fitted components is provided, suggesting that the limitations of conventional multinominal and random parameter logit models might also be pertinent to our dataset. However, the limitations inherent to semi-parametric and non-parametric models, and difficulty in their results generalization and transferability in particular, have to be borne in mind when the goal is to use the modelling results in general cost-benefit framework.

Acknowledgment

This research was funded by the Czech Ministry of Transport grant no. CG712-111-520: Quantification of external costs of transport in the Czech Republic. The views expressed in this paper are solely those of the authors.

References

Ben-Akiva, M., & Lerman, S. (1985). *Discrete choice analysis : theory and application to travel demand*. Cambridge Mass.: MIT Press.

Black, I. G. & Towriss, J. G. (1993). *Demand Effects of Travel Time Reliability*, Centre for Logistics and Transportation, Cranfield Institute of Technology, Great Britain.

Fosgerau, M. (2007). Using nonparametrics to specify a model to measure the value of travel time. *Transportation Research Part A: Policy and Practice, 41*(9), 842-856.

Louviere, J. J., Hensher, D. A. & Swait, J. D. (2000). *Stated choice methods : analysis and applications*. Cambridge, UK ; New York, NY, USA: Cambridge University Press.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp 105-142). Academic Press.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramjerdi F., Flügel S., Samstad H. & Killi M. (2010). Den norske verdsettingsstudien, Tid. *TØI* rapport *1053B*/*2010*, Oslo: Transportøkonomisk institut.

Small, K. A., Noland, R., Chu, X. & Lewis, D. (1999). Valuation of Travel-Time Savings and Predictability in Congested Conditions for Highway User-Cost Estimation, *NCHRP Report*, Vol. 431, Transportation Research Board, Washington: National Academy Press.

Small, K., & Verhoef, E. T. (2007). *The economics of urban transportation*. New York: Routledge.

Train, K. (2003). *Discrete choice methods with simulation*. New York: Cambridge University Press.

Wood, S. (2006). *Generalized additive models : an introduction with R*. Boca Raton: Chapman & Hall/CRC.

Contact

Vojtěch Máca
Charles University Environment Center
José Martího 2, Praha, Czech Republic
vojtech.maca@czp.cuni.cz

Jan Melichar
Charles University Environment Center
José Martího 2, Praha, Czech Republic
jan.melichar@czp.cuni.cz