Measuring errors with latent variables in transport models

Juan Manuel Lorenzo Varela[1], Maria Börjesson, Andrew Daly*

The Value of Travel Time (VTT) is fundamental in transport economics. In the last 30 years the best practice for VTT estimation has been to use Stated Choice (SC) data. However, there is now considerable evidence of reference dependence and gain-loss asymmetry in SC data, implying that such data do not reveal long-term preferences. This is a serious problem since the value of time is often applied in welfare analyses, where long-term stability of the preferences is a key assumption. A potential reason for the strong reference dependence found in SC data is the emphasis on a short-term reference point often used in SC data to reduce hypothetical bias. In the long-run there is no stable reference point.

An alternative to SC data is to use revealed preference data and a mode choice model to estimate the VTT. Observed behaviour has adapted to travel conditions and should thus be ruled by long-term preferences. Many countries collect NTS (national travel survey) data and spend considerable resources on making them representative, which is an argument for using them for VTT estimation. However, a key problem in the use of NTS data for VTT estimation is the measurement errors in the travel time and travel cost variables. Time and cost in NTS data is either self-reported or derived from a network assignment model.

In this paper we explore the errors in the self-reported and network-computed time and cost variables by treating travel time and travel cost as latent variables in the estimation of a mode choice model. We use Swedish NTS data, and a Transcad network to simulate travel time and cost with the state-of-practice method in large-scale modelling. We show how the magnitude of the errors in the input variables can be quantified, and explore the possibility of controlling for these errors to estimate the VTT on NTS data. We also explore the errors in the time and cost variables in a descriptive analysis, for instance with regard to rounding errors and driving costs.

We admit that we face a potential identification problem, i.e. that the random error in the choice model cannot easily be separated from error in the latent time and cost variables. In this case the assumption of the error structure in the choice model influences the estimated errors in the time and cost variables. We explore this issue by making sensitivity tests in the model formulation. However, given that we use a state-of-practice mode choice model, we argue that is it also relevant to explore the errors in the time and cost variables given this model. We use maximum likelihood measurement-error models, focussing on their application to mode choice models. To our knowledge, no previous study on large-scale transport models has explored the impacts of different model assumptions in error quantification in this way.

Previous studies have shown that regression models - including discrete choice models - are sensitive to errors in variables unless they are made explicit, hence parameter estimates are biased towards zero, an effect known as regression dilution. Perception, reporting (e.g. rounding) and modelling errors are just a few of the errors transport

[1] KTH Royal Institute of Technology, Department of Transport Science, Stockholm, Sweden
* Contact email address: jmlv@kth.se

models face; therefore, models that can account for these errors are of paramount importance to prevent biased estimates.

Based on previous work, we start by estimating a nested logit model, assuming no error in the input variables. Subsequent models use latent variables to quantify errors in the time and cost variables. We estimate different model specifications, some of which exploit specific model assumptions of the distributions for the latent variables, and show how this influences results. Moreover, we test how the results depend on the assumed error distribution of the time and cost variables.

We apply the 2005/06 Swedish National Survey for the Greater Stockholm Region. This provides 3485 observations, 1556 used PT, of which 46% used bus, 41% used metro and 13% used train. The remaining 1929 observations are divided between walk, bicycle, car driver and car passenger.

The results indicate that time and cost variables, normally used in mode choice models, whether reported or derived from networks, carry errors with them; hence, parameter estimates are diluted, and therefore biased. We also find that assumptions regarding the latent variable prior distribution affect parameter estimates, and that skewed distributions for time and cost variables, outperform the normal distribution. The goodness of fit of the assumed error distributions for time and cost variables was measured through the analysis of the measurement equation residuals, and we find that, ceteris paribus, multiplicative measurement-error models outperform additive ones.

The error quantification shows that residuals for cost variables exhibit much larger variance than do time variables. This suggests that cost parameters incur larger errors than time parameters.

In our data, an advantage for analysing time over cost variables is that two time indicators are available for the chosen alternatives – calculated travel time by the assignment software and self-reported time. Results show that whilst the time error variance for public transport modes is similar between the two indicators, we can observe large discrepancies for the car alternatives, the variance of the residuals for reported travel times being larger.

The resulting VTT from the different models are reported. Models not accounting for measurement errors yield higher values of time – between two and four times the values currently used in appraisal – than the models with latent variables, primarily due to higher cost parameter estimates. Results thus suggest that when the model do not account for errors in the travel cost variables, cost parameter estimates are diluted resulting in a too high VTT. Furthermore, VTT estimates from the final specification with latent values yield lower estimates than current VTT from SC data used in appraisal.