I T L S

WORKING PAPER
ITLS-WP-06-03

Selection Bias in Value of Travel Time Savings

By

Stefan L Mabit*

Visiting PhD student, Sept 04 – Jan 05
Centre for Traffic and Transport (CTT)
The Technical University of Denmark

March 2006

ISSN 1832-570X

INSTITUTE of TRANSPORT and LOGISTICS STUDIES

The Australian Key Centre in Transport and Logistics Management

The University of Sydney

Established under the Australian Research Council's Key Centre Program.

NUMBER: Working Paper ITLS-WP-06-03

TITLE: Selection Bias in Value of Travel Time Savings

ABSTRACT: In this paper we investigate the value of travel time savings. Estimation of this is in most cases based on samples using a specific mode. We use a mixed logit model to estimate the VTTS together with an auxiliary probit equation to capture the fact that the sample is non random. The results show that the probit equation in some cases gives extra information that can be used to improve the VTTS estimates from the mixed logit model. Hence this opens a way to investigate the possible selection bias in standard estimations of VTTS.

KEY WORDS: ***Discrete choice, value of travel time savings, mixed logit and selection bias.***

AUTHOR: Stefan L Mabit

CONTACT: Institute of Transport and Logistics Studies (C37)
An Australian Key Centre
The University of Sydney NSW 2006 Australia

Telephone: +61 9351 0071
Facsimile: +61 9351 0088
E-mail: itlsinfo@itls.usyd.edu.au
Internet: http://www.itls.usyd.edu.au

DATE: March 2006

1. Introduction

Concerning the estimation of The Value of Travel Time Savings(VTTS) there has emerged a commonly accepted framework derived from [DeSerpa, 1971] and a general theory on discrete choices. This framework has shown its use both with revealed preference data(RP) and stated preference data(SP). Advances concerning SP data has made this a preferred choice in recent European studies, see e.g. [Axhausen, 2004]. Since SP data makes it possible to ask any person within a population to make choices this approach underlines a question that also exist with RP data. If we want to estimate the average VTTS for a given population, which part of the population do we ask SP choices and how do we correct the results if we only give them to a subpopulation.

The above is best illustrated through an example. Suppose that our interest is the average VTTS in car for a population. Here we would choose a subpopulation and have them make SP choices. Three possible ways of choosing the subpopulation would be a random sample, a sample of people who has the possibility of using car to work or a sample of people always using car. But looking through the literature we have not been able to find discussions of the effect of choosing a given subsample[1] .

One way to investigate the effect of the sample is through a selection equation. Therefore we present a model with a selection equation and a VTTS equation. Estimations on data from a Danish VTTS study confirm that for some specifications there is significant correlation which corresponds to the fact that a standard estimation would suffer from selection bias.

In this paper we adopt the theory of DeSerpa and the theory on discrete choice as presented in [Ben-Akiva and Lerman, 1985]. Within the framework the VTTS can be found as the ratio between the time coefficient and the cost coefficient in the discrete choice model, see [Jara-Diaz, 2003]. The framework has over the last decade been enriched by the use of mixed models. This is explained very clearly in [Train, 2003]. One particular mixed model is the mixed logit model. It is a very flexible model, see [McFadden and Train, 2000], and practical in handling panel data, see [Revelt and Train, 1998]. A crucial issue is which distribution to use in the mixed logit model, since this implies a distribution on the VTTS, see [Hess et al., 2004], [Fosgerau, 2004] and [Hensher and Greene, 2003].

In transportation research the problem of selection bias has been recognized and studied in relation with choice based sampling. For choice based sampling the concern that a non random sample could lead to selection bias dates back to the late 70s. It was shown that standard estimation lead to inconsistent estimates, see [Manski and Lerman, 1977]. The problem in this context is that the object of study the choice affects the probability of being sampled. In the case of a VTTS SP-study the sample is based on mode choice which is based on VTTS. Hence it is reasonable to suspect that the sampling affects the outcome of the VTTS estimation.

*Centre for Traffic and Transport, Technical University of Denmark, slm@ctt.dtu.dk
[1]The only discussion found seem to be whether the sample shares some aggregated statistics of the population

It resembles the estimation of a wage equation in labor supply where the sample consist of people in a job. Since the accept of a job offer depends on peoples reservation wage one would suspect selection bias. In the field of labor supply this problem has been investigated through the use of a selection equation. The approach dates back to [Heckman, 1979]. He added a probit selection equation to the classical linear regression model. His approach has been enlarged to the case when both the selection and the main equations are discrete, see[Vella, 1992]. But in both cases the equations are connected through correlation in the additive error.

In this study interest is on VTTS hence we are interested in whether the sampling affects the estimation of coefficients in the model and not the additive error. Therefore the model here though resembling the model discussed by Vella is different since the correlation is captured as an interaction between the selection equation and the mixed parameters of the VTTS equation.

The remainder of this paper is organized as follows. In section 2 we present the model and estimation of the model. Section 3 contains a discussion on the data and section 4 discusses the model applied to data. The final section contains some concluding remarks.

2. Model

The model we present here consists of a Mixed logit model for panel data and a selection model to capture the way individuals were selected into the data set. In the next sections we present the two models and derive their simultaneous log likelihood.

2.1 Selection model

The selection model is a binary probit model. Let *n* denote a given individual. Then the binary outcome of the selection, Y_n, depends on a latent variable U_{1n}

$$Y_n = 1\{U_{1n} > 0\}, \tag{1}$$

where 1{} denotes the indicator function and U_{1n} is a latent variable determining the choice. Conditional on explanatory variables and coefficients the latent variable is the sum of a deterministic part and a random error term:

$$U_{1n} = \gamma' x_{1n} + u_{1n}, \tag{2}$$

where $u_1 \sim N(0,1)$ and x_{1n} denotes the explanatory variables including a constant. Based on the model we get that

$$P(Y_n = 1|\gamma, x_{1n}) = \Phi(\gamma' x_{1n}). \tag{3}$$

The expected residuals, $E(u_{1n}|Y_n = 1, x_{1n})$, have in many cases been used when sample selection is a concern, see e.g. [Heckman, 1979]. From the probit model it follows that:

$$E(u_{n1}|Y_n = 1, x_{1n}, \gamma) = \frac{\phi(\gamma' x_{1n})}{\Phi(\gamma' x_{1n})} = \lambda(\gamma' x_{1n}), \tag{4}$$

see [Wooldridge, 1999]. This is known as the Mills ratio. The use of a selection equation like the one described calls for a discussion of instruments. The variables in the selection equation not in the VTTS equation will act as instruments for the selection. Since a discussion like this depends on the specific application in mind we will postpone this discussion until section 3.

2.2 VTTS model

The VTTS model is a mixed logit model for panel data. In such a model one has to choose a distribution for the coefficients. This model assumes log normal coefficients i.e. log normal VTTS. The distribution ensures positive VTTS, see [Hess et al., 2004], and that VTTS has a well defined mean. Another reason to choose a log normal distribution is that on the same data set it has been shown to perform well in non parametric investigation of VTTS, see [Fosgerau, 2004]. The model parameterizes the VTTS with explanatory variables to allow for heterogeneity and heteroscedasticity.

The model applies to a panel of binary choices. In our notation each alternative only has two attributes. For each person with Y_n=1 we observe repeated SP choices:

$$Z_{nt} = 1\{U_{2nt} > 0\}, \tag{5}$$

$$Z_n = \{Z_{n1}, Z_{n2}, ..., Z_{nN_n}\} \tag{6}$$

The choice depends on the latent variable U_{2n}. The latent variable is decomposed into a random term and a systematic term conditional on random taste coefficients

$$U_{2nt} = V_{nt} + \varepsilon_{nt} = \alpha_n^C \Delta C_{nt} + \alpha_n^T \Delta T_{nt} + \varepsilon_{nt}, \tag{7}$$

where ε_{nt} is independent logistic and we have left out the alternative specific constants because the SP experiment is unlabeled. The $\Delta C_{nt}, \Delta T_{nt}$ refer to the difference in the cost attributes and time attributes between the two alternatives in a SP choice. Under the above assumptions we have

$$P(Z_{nt} = z_{nt}|Y_n = 1, V_{nt}) = \frac{e^{V_{nt} z_{nt}}}{1 + e^{V_{nt}}} \tag{8}$$

and

$$P(Z_n = z_n|Y_n = 1, V_n) = \prod_t \frac{e^{V_{nt} z_{nt}}}{1 + e^{V_{nt}}}. \tag{9}$$

Where V_n denotes the vector(V_{nt}). The parameters α_n^C, α_n^T are chosen so that the VTTS is log normal distributed as mentioned above, i.e. the fraction $\frac{\alpha_n^T}{\alpha_n^C}$ follows a log normal distribution. This can be done in different ways. We will use the following specifications

$$\alpha_n^C = e^{\beta' x_{2n} + \sigma_2 u_{2n}} e^{\beta_T + \sigma_3 u_{3n}}, \text{ and } \alpha_n^T = e^{\beta_T + \sigma_3 u_{3n}}, \tag{10}$$

where e.g. $\alpha_n^T = e^{\beta_T + \sigma_3 u_{n3}}$ signifies that α_n^T follows a log normal distribution derived from a $N(\beta_T, \sigma_3^2)$. We allow the cost coefficient to be parameterized by explanatory variables, x_{2n}. The specification is what has been named estimation in WTP-space, see [Train and Weekes, 2005]. The use of WTP-space as opposed to the classical preference space is still limited. The reason for using it here is that preliminary investigations gave better models and also that [Fosgerau, 2005] shows that it performs well when compared with non-parametric estimations. Another reason for the WTP-specification is that correlation is modeled directly whereas it will only be indirectly through correlation with the cost coefficient in preference space. From the two specifications we get:

$$\text{VTTS} = e^{-(\beta' x_{2n} + \sigma_2 u_{2n})}. \tag{11}$$

Which is log normal distributed.

The model for selection from section 2.1 and the model for VTTS estimation above allow for interaction through correlation of the different normal distributions. This gives a model where it is possible to test whether the selection equation influences the VTTS estimation. Assuming that u_1 and u_2 follow a joint normal distribution we can use a Choleski factorization and write

$$u_{n1} = v_{n1},\ u_{n2} = s_1 v_{n1} + s_2 v_{n2} \text{ and } u_{n3} = s_3 v_{n3} \tag{12}$$

where v_i are iid normal with mean zero and variance one.

Now we can derive a simultaneous likelihood for the selection and the SP-choices. We observe a vector of choices Z_n when Y_n=1. We condition on $x_n, \Delta T, \Delta C,$ but leave this out of the notation.

$$\begin{aligned} P(Z_n = z_n, Y = 1) &= E\left(P\left(Z_n = z_n, Y_n = 1 | v_n\right)\right) \\ &= E\left(P\left(Z_n = z_n | Y_n = 1, v_n\right) P\left(Y_n = 1 | v_n\right)\right) \\ &= E\left(\prod_{t=1}^{N_n} \frac{e^{V_{nt} z_{nt}}}{(1 + e^{V_{nt}})} 1\left\{v_{n1} > -\gamma' x_n\right\}\right) \\ &= \int\int\int_{-\gamma x_n}^{\infty} \prod_{t=1}^{N_n} \frac{e^{V_{nt} z_{nt}}}{(1 + e^{V_{nt}})} \phi\left(v_1\right) \phi\left(v_2\right) \phi\left(v_3\right) \partial v \end{aligned} \tag{13-16}$$

where we parameterize by $v = (v_1, v_2, v_3)$ and

$$\begin{aligned} V_{nt} &= e^{\beta_C' x_{2n} + u_2} e^{\beta_T + u_3} \Delta C_{nt} + e^{\beta_T + u_3} \Delta T_{nt} \\ &= e^{\beta_C' x_{2n} + s_1 v_1 + s_2 v_2} e^{\beta_T + s_3 v_3} \Delta C_{nt} + e^{\beta_T + s_3 v_3} \Delta T_{nt}. \end{aligned} \tag{17-18}$$

2.3 Estimation

We have in the above calculations only focused on the case Y=1. A equivalent model for a panel of binary choices for individuals with Y=0 can be deduced in the same way. The model can be estimated by classical methods such as simulated maximum likelihood(SML), see [Train, 2003]. Here we will use a simultaneous approach so we estimate the two equations simultaneously using SML. One could use a sequential approach involving the mills ratios to test the hypothesis of correlation. But in case of a rejection of the null hypothesis the estimates would be biased in our case in contrast with the Heckman model where a correction of the standard errors is enough.

In the model VTTS is assumed log normally distributed. To evaluate the model and compare it to the standard model we evaluate the mean VTTS. This can be done in two ways either by averaging over the sample or by choosing a representative individual. The first is useful in applications but for the purpose at hand where we compare two models the second approach is more appropriate since model differences are not confused with sample characteristics. Therefore we evaluate the means using an individual with mean socioeconomic variables i.e. $\overline{x}_i = \frac{1}{N}\sum \mathrm{x_i}$ and $\overline{x} = (\overline{x}_i)$.

We have :

$$\frac{\alpha_T}{\alpha_C} = e^{-\beta'_C x_2 - s_1 v_1 - s_2 v_2} \tag{19}$$

where v_1, v_2 are independent normal. Therefore we get

$$VTTS = E(\frac{\alpha_T}{\alpha_C}|\overline{x}) = e^{-\beta'_C \overline{x}_2 + \frac{1}{2}(s_1^2 + s_2^2)}. \tag{20}$$

This expression depends on β and s. We could use the point estimates for β and s, but to be true to the fact that they are estimates, confer [Hensher and Greene, 2003], it is more correct to use the estimated asymptotic distribution of (β,s). To do this we draw M time (β_m, s_m) and with these calculate:

$$VTTS_m = e^{-\beta'_{mC} \overline{x}_2 + \frac{1}{2}(s_{m1}^2 + s_{m2}^2)}, \tag{21}$$

Then we use the average

$$VTTS = \frac{1}{M}\sum_m VTTS_m \tag{22}$$

as the resulting VTTS estimate from the model. As an estimate of the variation in the VTTS estimate we calculate

$$std = (\frac{1}{M}\sum_m (VTTS_m - VTTS)^2)^{\frac{1}{2}} \tag{23}$$

and report this as std. In the same way we can estimate the mean conditional on the individual $\overline{x}$ being a car user i.e.

$$E(\frac{\alpha_T}{\alpha_C}|\overline{x}, Y=1) = \frac{\Phi(\gamma'\overline{x}_1 + s_1)}{\Phi(\gamma'\overline{x}_1)} e^{-\beta'_C\overline{x}_2 + \frac{1}{2}(s_1^2+s_2^2)}. \quad (24)$$

Conditional on being a public transport user the expression becomes:

$$E(\frac{\alpha_T}{\alpha_C}|\overline{x}, Y=0) = \frac{\Phi(-\gamma'\overline{x}_1 - s_1)}{\Phi(-\gamma'\overline{x}_1)} e^{-\beta'_C\overline{x}_2 + \frac{1}{2}(s_1^2+s_2^2)}. \quad (25)$$

If we had estimated (β_s, s_s) from a standard model without selection we could do the same calculation using

$$E(\frac{\alpha_T^s}{\alpha_C^s}|\overline{x}) = e^{-\beta'_{sC}\overline{x}_2 + \frac{1}{2}(s_s^2)}. \quad (26)$$

3. Data

The data is taken from the 2004 Danish VTTS study, see [Burge, 2004]. In this paper we will only investigate the commuters using car or public transport(PT). The data consists of 1425 individuals. Each individual was asked 9 SP choices in an unlabeled experiment referring to a current commute trip. Every choice was a binary choice where the alternatives were only described by travel time and cost. The SP choices included a check question i.e. a choice where one alternative is slower and more expensive. A total of 177 persons chose this dominated alternative. Since we cannot be sure that these people understood the SP task we take them out of the sample[2] .

Of the remaining 1248 individuals 3 had unrealistic reported travel times, 5 had unrealistically large travel costs, 24 had unrealistic travel speeds and 1 didn't complete all of his choices. This leaves us with 1216 individuals of which 739 used public transport and 477 car. For the discrete variables the individuals in the sample have the characteristics reported in Table 1. The variables in the table have the following definitions.

Table 1: Descriptive statistics of the data

variable	per	per	PT per	PT per	Car per	Car per
area/area2	37.9	34.2	47.8	24.5	22.6	49.5
cars/nocar	16.5	21.5	8.3	33.6	29.4	2.7
carsin/not	8.8	91.2	5.3	94.7	14.3	85.7
grp2/grp3	6	5.9	2.0	4.3	12.2	8.4
hinc/noinc	32.2	5.5	33.7	5.0	29.8	6.3
lic/nolic	91.0	9.1	85.5	14.5	99.4	0.6
lug/nolug	11.7	88.3	10.0	90.0	14.3	85.7
male/female	53.4	46.6	49.8	50.2	58.9	41.1
tripf/not	41.9	58.1	53.2	46.8	24.3	75.7
weekend	4.7	95.3	2.8	97.2	7.6	92.5
work home	80.8	19.2	81.9	18.1	79.3	20.8
occupation	98.1	1.9	97.2	2.8	99.6	0.4

[2]Whether this is good practice is questionable. Seven percent seems a lot.

The variable area is an indicator that the persons home is in Copenhagen and area2 an indicator that it is in a small town with less than 10000 inhabitants. The variable cars is an indicator of more than one car in the household, while nocar is an indicator for no car. The variable carsin is an indicator for living in a single adult household with one car. The variable grp2 indicates if traveling with household member and grp3 if traveling with non household member. The income variable noinc is an indicator for people with unknown income, while hinc is an indicator for people with high household income i.e. more than 600.000 DKK pr. year[3] . The variable lic is an indicator for having a licence and the variable lug is an indicator for traveling with large luggage. The variable male/female is self explanatory. The variable tripf indicates that the travel takes place less often than daily. The variable weekend indicates that the travel takes place on the weekend. The variable occupation indicates if the individual has an occupation as a wage earner or self employed as opposed to apprentice. The variable work home indicates wage earners and self employed working at home less than once a week.

Table 2: More descriptive statistics

variable	mean	std.der.	min	max	mean(pt)	mean(car)
age	42.8	11.2	16	73	41.4	45.0
logdis	3.02	1.13	0	6.40	2.98	3.08
loginc	1.26	0.51	0	2.40	1.25	1.27
logtime(min)	3.22	0.80	1.10	5.86	3.25	3.17

The characteristics of the continues variables can be seen in Table 2. Here age is the age, logdis is the log of the distance between origin and destination. If distance was set to zero logdis is set to zero. The variable loginc is the log of personal income for the people with reported income(inc is not continues, but discrete with 11 levels). The variable logtime is the log of reported travel time. It is worth remarking that car users are older, travel longer distances in shorter time and income is the same in the two segments.

As mentioned above the data consists of SP-games on time and cost in current mode. For each individual there are 9 SP observations. We use only 8 of them because one was a check question. The reasons for not using the check question are that the SP design is balanced without the check question and also that the information given by the dominated choice is uninformative since we base our model on the theory of De Serpa which results in non-negative VTTS.To estimate the model the alternatives have been arranged such that alternative 1 is the fastest. With this rearrangement the differences of the attributes of the SP choices can be seen in Table 3.

[3]For conversion 1 euro is around 7.5 dkk

Table 3: Statistics on the attributes

variable	mean	std.der.	min	max	mean(pt)	mean(car)
diffT	-7.24	6.9	-60	-1	-7.77	-6.42
diffC	6.94	11.2	0.5	200	7.11	6.68

Where diffT denotes the time attribute of the first alternative minus the time attribute of the second etc.. Each alternative in the SP choices concerning car also included the attribute congested time. Since this attribute in all choices had a fixed ratio to the total time depending on reported congestion we choose to use only total travel time and include the congestion ratio as an explanatory variable. This approach was also used in [Fosgerau, 2004].

3.1 Instruments

An important concern with a model using two equations is which variables to include in the VTTS equation as explanatory variables and which to include in the selection equation as instruments. As explanatory variables in the VTTS equation we choose to include income and time, since this is supported by theory. We also choose to include age and sex since the causality between these variables and VTTS is clear. The inclusion of these two variables has essentially the same purpose as segmentation. For the remaining variables we do not have theory to support their inclusion in the VTTS equation and since causality between them and VTTS is not clear we would run the risk of endogenous variables if we included them. Therefore they will function as instruments in the selection equation.

4. Results

Using the data we have estimated the model without selection as a separate probit and mixed logit model. The estimations were done using a program written in Ox, see [Doornik, 2001]. The program used SML to maximize the likelihood function in 13. In Table 4 we report the results for each of the two segments using this standard model.

Table 4: Estimation results for independent models space

parameter	estimate	t-value	estimate	t-value
Segment	car		pt	
LL	-509.6			
constant	-3.12	(-4.28)		
γ_{age}	0.12	(2.78)		
γ_{area}	-0.40	(-3.41)		
γ_{area_2}	0.33	(2.90)		
γ_{carno}	-1.42	(-8.15)		
γ_{cars}	0.87	(7.01)		
γ_{carsin}	0.54	(3.55)		
γ_{child}	-0.27	(-1.92)*		
γ_{logdis}	-0.15	(-3.69)		
γ_{grp2}	1.53	(7.23)		
γ_{grp3}	0.59	(3.11)		
γ_{hinc}	-0.24	(-2.30)		
γ_{lic}	1.71	(4.61)		
γ_{lug}	0.44	(3.05)		
γ_{sex}	-0.32	(-3.43)		
γ_{tripf}	-1.08	(-11.0)		
$\gamma_{weekend}$	0.65	(2.77)		
γ_{occup}	1.88	(3.14)		
$\gamma_{workhome}$	-0.28	(-2.31)		
LL	-2028.4		-2712.6	
halton draws	1000		1000	
β_T	-1.22	(-10.5)	-0.79	(-10.1)
β_0	1.33	(5.26)	2.32	(13.0)
β_{age}	0.30	(8.96)	0.15	(3.75)
β_{inc}	-0.43	(-5.21)	-0.82	(-10.2)
β_{noinc}	-0.91	(-6.32)	-1.26	(-6.74)
β_{sex}	0.21	(3.74)	-0.001	(-0.01)*
β_{time}	-0.50	(-9.65)	-0.36	(-6.04)
β_{cong}	-1.87	(-6.74)	no	no
s_2	1.05	(20.9)	1.02	(23.1)
s_3	1.66	(9.21)	1.22	(11.5)

The parameters refer to the notation introduced in section 3 e.g. γ_{area} multiplies the indicator for area. There are a few exceptions they are: γ_{sex} that multiplies an indicator for female. Remember that the selection model only uses socio-demographic variables so the parameters are differences in the effect of a variable on utility of car and pt.

The parameters in the selection equation are all significant. The result is the same for PT in this estimation hence the estimates are not reported. For some of the parameters it is hard to have an expectation on the sign beforehand. For the ones with expected signs e.g. cars, nocar, age, sex, area, lic we get the expected signs.

For the parameters in the VTTS equation we get similar results for the two segments. The only exception is sex which is significant with the expected sign for car i.e. lower VTTS for females, but which is insignificant with the opposite sign for public transport. Also it is worth noticing that the effect of income is higher in PT while age and travel time has a higher effect on VTTS in car. From the specification we get that VTTS rises with income, time, congestion and falls with age, sex.

The results for both segments using the model with correlation are reported in Table 5. The parameters refer to the same variables as in Table 4, except for s_1 which is the coefficient on correlation. Concerning the parameters from the selection equation they are all significant. For both segments we get the same signs and similar estimates as for the model without correlation.

Table 5: Estimation results for model with correlation

parameter	estimate	t-value	estimate	t-value
Segment	car		pt	
halton draws	1000		1000	
LL	-2537.2		-3221.1	
constant	-3.02	(-5.30)	-3.05	(-5.30)
γ_{age}	0.10	(3.02)	0.11	(2.44)
γ_{area}	-0.44	(-4.49)	-0.42	(-3.93)
γ_{area_2}	0.31	(3.00)	0.32	(3.10)
γ_{carno}	-1.41	(-12.3)	-1.40	(-8.55)
γ_{cars}	0.90	(8.05)	0.87	(8.68)
γ_{carsin}	0.56	(4.18)	0.51	(3.51)
γ_{child}	-0.28	(-2.63)	-0.28	(-2.38)
γ_{logdis}	-0.16	(-5.84)	-0.17	(-4.78)
γ_{grp2}	1.56	(8.02)	1.55	(11.1)
γ_{grp3}	0.57	(3.72)	0.62	(3.66)
γ_{hinc}	-0.26	(-2.80)	-0.24	(-2.76)
γ_{lic}	1.65	(5.88)	1.67	(5.63)
γ_{lug}	0.44	(3.95)	0.46	(2.62)
γ_{sex}	-0.32	(-3.99)	-0.33	(-4.03)
γ_{tripf}	-1.10	(-13.0)	-1.09	(-12.5)
$\gamma_{weekend}$	0.63	(2.71)	0.67	(2.84)
γ_{occup}	1.95	(4.15)	1.92	(4.42)
$\gamma_{workhome}$	-0.27	(-2.68)	-0.27	(-2.78)
β_T	-1.21	(-10.4)	-0.79	(-10.1)
β_0	1.36	(6.71)	2.37	(14.5)
β_{age}	0.28	(9.46)	0.12	(5.98)
β_{inc}	-0.39	(-5.77)	-0.81	(-9.45)
β_{noinc}	-0.68	(-5.34)	-1.30	(-7.76)
β_{sex}	0.21	(3.26)	-0.04	(-0.73)*
β_{time}	-0.49	(-12.0)	-0.35	(-9.02)
β_{cong}	-1.89	(-6.67)	no	no
s_1	-0.10	(-3.70)	-0.11	(-3.38)
s_2	1.05	(23.9)	1.02	(28.6)
s_3	1.69	(9.39)	1.23	(11.5)

For the parameters in the VTTS equation we get similar result for the two segments with two exceptions: the first is the same as the model above with the variable sex. For the car segment the signs on all parameters are the expected ones[4] i.e. VTTS rises with

[4]remember that the parametrization is for the inverse VTTS

income, time, congestion and falls with age, sex. Also we see that the correlation is has a negative sign i.e. the fact that people choose car explains part of their higher VTTS.

In the pt segment sex is insignificant. Again the signs on all significant parameters are the expected ones i.e. VTTS rises with income, time, and falls with age. Again we see a higher effect of income and lower of travel time and age on VTTS in PT. We see that the correlation has the same sign as in the car segment. The reason why the signs are expected to be the same for the two segments follows from fact that we estimate the same selection equation for the two segments. Therefore the truncations in the selection equation are different for the two segments, hence the same sign on s_1 indicates opposite effects of the correlation in the two segments.

4.1 VTTS estimation

The estimated VTTS for the average individual in the sample is calculated with formula (20) and the estimated parameters from Table 5. The results are seen in Table 6 with std. dev. in parenthesis calculated using equation 23. It is evaluated with congestion zero to compare across segments.

Table 6: VTTS for average individual in sample in DKK pr. Min

Model/segment	car	pt
VTTS	1.00(0.03)	0.89(0.04)

We also calculate the VTTS for people in a specific mode using 24 and 25, see 7.

Table 7: VTTS for average individual in sample in DKK pr. min conditional on mode

Model/segment	car	pt
VTTS	1.15(0.06)	0.82(0.04)

For comparison we also calculated the VTTS for people in a specific mode using the standard specification without selection from 26, see 8.

Table 8: VTTS for average individual in subsample in DKK pr. min conditional on mode

Model/segment	car	pt
VTTS	1.07(0.06)	0.85(0.04)

We can draw two immediate conclusions: In the standard case one would conclude that the VTTS in car is higher than in PT. This conclusion is not so obvious when taking selection into account. The second conclusion is that the bias from selection is neutralized by bias in parameter estimates so that mode specific results are much more different with the selection model than with the standard model.

5. Conclusion/Remarks

In this paper we have estimated the VTTS taking into consideration that the sample might not be random in the population. To estimate the VTTS we have included a selection equation. Both with and without the selection equation we obtain reasonable VTTS distributions having significant parameter estimates with expected signs.

For both segments we calculate the VTTS using three different formulas. They show that taking selection into account for the two segments changes the expected VTTS toward one another. The calculations using the standard model show that these values resemble the values from the general model. But the fact that bias in estimates correct for selection bias should not be seen as a reason for not taking selection bias seriously but more as a reminder that hopefully the errors are not to large in cost benefit analysis when using the standard approach. An important conclusion is that whereas the standard estimates support a hypothesis that VTTS in car is higher than VTTS in public transport it is not an obvious conclusion from the estimates corrected for selection.

The main conclusion from this paper is that the sample selection effects VTTS estimation and that the effect can alter the final output from the model. A second conclusion is that selection bias should be looked at more seriously in the transportation field and others especially now that researches have moved away from the simple MNL model. So more research is needed to investigate if it is just a sample and/or model specific selection effect we have found.

References

[Axhausen, 2004] Axhausen, K. e. a. (2004). Swiss value of travel time savings. Technical report, IVT. IVT.

[Ben-Akiva and Lerman, 1985] Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis*. MIT Press, Cambridge, MA, USA.

[Burge, 2004] Burge, P., R. C. (2004). Dativ: Sp design: Proposed approach for pilot survey. Technical report, TetraPlan in cooperation with RAND Europe and Gallup A/S. TetraPlan.

[DeSerpa, 1971] DeSerpa, A. (1971). A theory of the economics of time. *The Economic Journal*, 81:828–846.

[Doornik, 2001] Doornik, J. (2001). *Ox: An Object-Oriented Matrix Language*. Timberlake Consultants Press, London.

[Fosgerau, 2004] Fosgerau, M. (2004). Investigating the distri-bution of the value of travel time savings. *working paper*.

[Fosgerau, 2005] Fosgerau, M. (2005). Specification of a model to measure the value of travel time savings from binomial data. *Presented at ERSA, Amsterdam*.

[Heckman, 1979] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–162.

[Hensher and Greene, 2003] Hensher, D. A. and Greene, W. H. (2003). The mixed logit model: The state of practice. *Transportation*, 30:133–176.

[Hess et al., 2004] Hess, S., Bierlaire, M., and Polak, J. W. (2004). Estimation of value-of-time using mixed logit models.

[Jara-Diaz, 2003] Jara-Diaz, S. R. & Guevara, C. A. (2003). Behind the subjective value of travel time savings: The perception of work, leisure and travel from a joint mode choice-activity model. *Journal of Transport Economics and Policy*, 37:29–46.

[Manski and Lerman, 1977] Manski, C. and Lerman, S. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, 45 (8):1977–1988.

[McFadden and Train, 2000] McFadden, D. and Train, K. E. (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics*, 15 (5):447–70.

[Revelt and Train, 1998] Revelt, D. and Train, K. (1998). Mixed logit with repeated choice: Household's choices of appliance efficiency level. *The Review of Economics and Statistics*, 80(4).

[Train and Weekes, 2005] Train, K. and Weekes, M. (2005). *Discrete Choice Models in Preference Space and Willingness-to-Pay Space*, pages 1–16 in R. Scarpa and Alberini(eds) Application of simulation methods in environmental and resource economics. Springer.

[Train, 2003] Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press, New York, NY, USA.

[Vella, 1992] Vella, F. (1992). Simple tests for sample selection bias in censored and discrete choice models. *Journal of Applied Econometrics*, 7 (4):413–421.

[Wooldridge, 1999] Wooldridge (1999). *Econometrics and Panel Data*. Chapman & Hall, NY, USA.