A latent variable approach to dealing with missing or inaccurately measured variables: the case of income

Stephane Hess* Nobuhiro Sanko† Jeff Dumont‡ Andrew Daly§

Abstract

It is clear to most choice modellers that variables characterising alternatives and decision makers are potentially affected by a number of important problems. These include measurement error, correlation with other unobserved factors, systematic errors such as lying by respondents and missing values. Standard approaches are to ignore the first two, to assume that the third does not happen and either to remove responses with the fourth problem or to impute missing values from observations of other respondents. Income is a key example of such a variable, yet is of immense importance in many models. Issues with income arise as measurement errors in categorically captured income, correlation with unobserved variables such as accessibility to specific transport modes, systematic over- or under-statement of income and missing income values for those who refused to answer. The present paper illustrates how an analyst can deal with these issues by replacing reported income with a *latent* income variable in a choice model, which at the same time is also used to explain the *stated* income in a measurement model. In comparison with using stated income directly, this deals with at least some of the measurement error and bias issues. In comparison with using imputation of missing values, it draws not just on data on stated income for those respondents without missing information, but the simultaneous estimation with the choice model means that the observed choices also inform the latent income variable. Furthermore, unlike approaches relying on stated income or on imputed values, the method is directly applicable for forecasting. Two empirical applications using stated and revealed preference data illustrate the good performance of the method in practice.

Keywords: latent variables; missing income; discrete choice; random heterogeneity

*Institute for Transport Studies, University of Leeds, s.hess@its.leeds.ac.uk

†Graduate School of Business Administration, Kobe University and Institute for Transport Studies, University of Leeds, sanko@kobe-u.ac.jp

‡Institute for Transport Studies, University of Leeds and RSG, jeff.dumont@rsginc.com

§Institute for Transport Studies, University of Leeds and RAND Europe, daly@rand.org

1 Introduction

Latent variable (or hybrid) choice models are becoming increasingly popular in a number of disciplines, including transport (see e.g. Ben-Akiva et al., 1999, 2002a,b; Ashok et al., 2002; Bolduc et al., 2005). The models are used primarily for accommodating attitudes and perceptions, but have also been used to accommodate other behavioural phenomena such as the formation of plans leading to choices (Choudhury et al., 2010).

In the present paper, we explore how latent variable models can be used to deal with errors, biases or missing values in key variables. In this context, the paper builds on recent work by Walker et al. (2010) and Brey and Walker (2011). The method used also builds on the recognition that imputed values contain additional error, previously treated analytically by Daly and Zachary (1977) or by simulation as 'multiple imputation' by Brownstone and Steimetz (2005). We specifically deal with the case of income, which is clearly a key variable in transport demand models, and one which is regularly affected by issues with missing data, while the fact that income is generally captured as a categorical variable leads to measurement error[1]. Previous work by Bhat (1994) has looked at imputing a continuous income variable from such data. The difference in our work is that we use joint estimation with the choice model, meaning that the latent income variable is informed not just by the observed income levels (i.e. non-missing) but also by the choice behaviour in the data. Alongside presenting a general model specification, we produce empirical results from two case studies, one on stated preference (SP) data, and one on revealed preference (RP) data, both of them illustrating the good performance of the approach in practice.

The remainder of this paper is organised as follows. The following section discusses modelling methodology. This is followed in Section 3 and 4 by the two separate empirical examples, and Section 5 summarises the findings and presents our conclusions.

2 Methodology

Let the utility for alternative i in choice situation t for respondent n be given by $U_{int} = V_{int} + \varepsilon_{int}$, where the deterministic component V_{int} is a function of measured attributes x_{int}, measured or reported socio-demographic characteristics z_n, and estimated sensitivities β_n. The remaining random component of utility

[1] It has also been brought to our attention that the specific case of missing income is used as an example in the documentation of Biogeme (Bierlaire, 2003, 2005).

is defined as ε_{int}. A general specification is given by $V_{int} = f(\beta_n, x_{int}, z_n)$. There is scope for a number of possible problems for x_{int} and z_n, including measurement error, missing values, correlation with other unobserved factors, and systematic respondent caused error, for example in the form of under or over-reporting. It is commonly accepted that measurement error and missing values are likely to lead to higher error in our model, but that it should not cause bias in parameter estimates. This is however potentially not the case if, as is probable, measurement error or non-reporting are correlated with the values - for example, we typically have wider income bands for higher incomes and people with low and high incomes are believed to report less often than those with moderate incomes. Correlation with other unobserved factors potentially causes endogeneity bias, while systematic error could also lead to biased estimates. The standard approaches used in practice to deal with the problems listed above vary substantially.

Measurement error is often ignored in practice, with an assumption that its effects are captured in the error term of the model. However, with the growing reliance on random coefficients models, there is also a risk that error in measured attributes is captured in the form of taste heterogeneity, and this could result in biased estimates of taste heterogeneity.

Missing values commonly arise for income and a number of other socio-demographic variables, as well as for some explanatory variables in RP data. A common approach is to remove affected respondents from the data, which obviously leads to an undesirable reduction in sample size, and might also make the resulting dataset less representative of the real population. An alternative is to attempt to impute the concerned attribute for those respondents with missing information, a process that essentially links the values for those respondents where the attribute is observed to other measured attributes (e.g. income linked to age) and then uses that relationship to infer the value for those respondents with missing data. A key limitation of imputation is that it assumes that the relationship between the affected variable and the various other attributes used as explanators is the same across those respondents who report values and those who do not. Imputation is also only informed by the observed values for this variable for other respondents, and not for example by the observed choice behaviour of respondents with missing data. Furthermore, when a value is imputed, it comes with imputation error and this needs to be taken into account in estimation to avoid biasing (towards zero) of the relevant coefficient.

Correlation with other unobserved factors can cause bias, and while the above listed work on latent attitudes and other latent factors has made headway in this area, many studies still ignore the potential risk of such correlation, especially when it concerns explanatory variables in RP data.

Systematic errors are in some ways the greatest concern. If a respondent purposefully misrepresents reality, for example by over or under-stating key variables that are used in model estimation, then this is likely to have a detrimental effect on model results. Income may well be the most likely attribute to be affected by this problem.

We now proceed with the description of a simple latent variable approach for dealing with these issues. Let us assume that a given element in z_n is subject to the above issues, say z_{nk}. In this paper, we focus on an element of z_n; a corresponding approach can be used for elements in x_{int} but indicators might be required for each t. We now use a latent variable $\alpha_{nk} = \gamma' z_n^* + \eta_n$, where z_n^* is a subset of z_n which does not include z_{nk}, and where η_n follows a standard Normal distribution. Inside our choice model, we now replace z_{nk} by $\tau_k \alpha_{nk}$, and possibly with additional function transformations such as $e^{\alpha_{nk}}$ to ensure positive signs for parameters.

We then have that:

$$PC_n\left(C_n \mid x_n, z_n^*, \beta_n, \alpha_{nk}, \tau_k\right) = \prod_{t=1}^{T} PC_{nt}\left(c_{nt} \mid x_{nt}, z_n^*, \beta_n, \alpha_{nk}, \tau_k\right) \quad (1)$$

gives the probability of the observed sequence of choices for respondent n, where c_{nt} is chosen in task t. Typically, PC_{nt} will be of logit form, and is a function of observed attributes x_{nt} and z_n^*, estimated parameters β_n and τ_k as well as a specific realisation of the latent variable α_{nk}, where in practice, marginal utility coefficients β_n would not be individual specific in estimation though they may follow a distribution of a pre-specified shape.

Thus far, our specification does not make use of any additional information, and simply replaces z_{nk} by a construct composed of a deterministic component and a random component, where this deterministic component would be informed only by the choices observed in the data. This is in contrast with imputation, where the value is informed only by the *correctly* observed values for z_k for all other respondents, and not by any respondent's choice behaviour.

In our case, additional model components are now used to help *inform* the role of the latent variable. At a bare minimum, we would have a single indicator (I) for each latent variable, which would be given by the original value for z_{nk}. We would then explain the observed values through a measurement model, where the contribution by person n is given by the probability of the observed indicator (PI_n):

$$PI_n\left(z_{nk}\right) = g\left(\alpha_{nk}, \Omega_{Ik}\right), \qquad \text{if } z_{nk} \text{ is observed/reported} \quad (2)$$

$$PI_n\left(z_{nk}\right) = 1, \qquad \text{if } z_{nk} \text{ is missing,} \quad (3)$$

where Ω_{Ik} is a vector of parameters, and where Equation 3 ensures that missing observations do not contribute to the estimation of the latent variable.

Both model components (PC_n and PI_n) are a function of a specific realisation of the latent variable, and integration over the random component in α_{nk} is thus needed. We use a simultaneous specification, where the contribution by respondent n to the overall likelihood is given by:

$$L_n = \int_{\alpha_{nk}} PC_n\left(x_n, z_n, \beta_n, \alpha_{nk}, \tau_k\right) PI_n\left(z_{nk}\right) \phi\left(\alpha_{nk}\right) \mathrm{d}\alpha_{nk}, \tag{4}$$

where ϕ is the normal density function.

As already mentioned in the introduction, the specific focus of this paper is on income, albeit that the above framework is readily applicable to other variables. Income is a key variable in many choice models, mainly for explaining heterogeneity in cost sensitivity, and it also serves as a proxy for numerous other effects. Income is also almost uniquely subject to all of the issues mentioned above. Firstly, measurement error occurs, for example, because income is often captured in categories rather than continuous variables, meaning that the *observed* values are only an approximation of the *true* values. Secondly, missing values occur as many respondents refuse to give income, an issue compounded by the fact that they are often not a random sample. Thirdly, correlation with unobserved preferences (ε) occurs as stated income is likely to be strongly correlated with other factors that affect behaviour. Finally, systematic error may occur as respondents may purposely overstate (or understate) their income, or be truly unsure about it.

3 First case study

The data for the first case study comes from a stated choice survey for intra-mode commuter choices, using rail or bus (see Hess et al., 2012, for a full description). Respondents were faced with ten tasks each involving the choice between three alternatives, of which the first was a reference trip, with attributes held invariant across tasks. Alternatives were described by travel time, fare, the rate of crowding (0 to 1), the rate of delays (0 to 1), the average delay across delayed trips, and the availability of a free text message (sms) delay information service. A sample of 368 respondents was obtained from an internet panel, leading to $3,680$ observations in the data. Income was captured in 9 separate categories, with 12.5% missing from the final sample.

As a first step, Table 1 presents the estimates from a simple Multinomial Logit (MNL) model estimated on this dataset, with a linear in attributes spec-

Table 1: Estimation results for a simple MNL model for first case study

log-likelihood: -3,700.43

	est.	rob. t-rat.
δ_1	0.20262	3.33
δ_2	0.13135	3.01
β_{tt}	-0.03652	-7.72
β_{fare}	-1.0126	-4.37
β_{crowding}	-1.6805	-7.25
$\beta_{\text{rate of delays}}$	-1.7777	-7.71
$\beta_{\text{av. delay}}$	-0.03441	-4.87
$\beta_{\text{delay sms}}$	0.2705	4.48

ification[2], and with constants for the first two alternatives (δ_1 and δ_2). All parameters are statistically significant and of the expected sign.

As a next step, Table 2 presents the results for a MNL model with separate fare coefficients for the nine income groups, as well as additional coefficients for respondents who are unsure about their income and respondents who refuse to provide income. This segmentation of the cost coefficient leads to statistically significant gains in model fit (64.65 units for 10 additional parameters), but produces counter-intuitive results for income groups 4, 5 and 9, while otherwise, there is general trend of decreasing cost sensitivity as income increases.

Given the observation of a general trend in cost sensitivity (with some imperfections), we adopted a continuous specification in the next set of models. Specifically, we replace β_{fare} by $\beta_{\text{fare}} inc_n^{\lambda_{inc}}$ where inc_n is an approximation to the continuous income for respondent n, using category mid-points, while λ_{inc} is an estimated income elasticity. We use this specification in a Mixed Multinomial Logit (MMNL) model which makes use of lognormal distributions for all coefficients, and where a separate cost coefficient (without income interaction) is used for respondents with no income information. The results for this model are presented in Table 3. We observe very significant improvements in model fit over the base MNL model by 707.78 units for 9 additional parameters. The MMNL model also retrieves significant patterns of heterogeneity across respondents for all six attributes, and an income elasticity of -0.253.

Finally, we turn our attention to the hybrid choice model. The log-likelihood for this model is $-3,583.58$ which relates to the combined choice model and

[2]The units for travel time and average delays are in minutes, those for fare are in £, the rates of crowding and delays are expressed from 0 to 1 and a simple dummy is used for the delay information sms.

Table 2: Estimation results with separate cost coefficients by income group for first case study

log-likelihood: -3,635.78

	est.	rob. t-rat.		est.	rob. t-rat.
δ_1	0.2300	3.84	$\beta_{\text{fare, inc}< £5K}$	-2.9262	-2.93
δ_2	0.1426	3.20	$\beta_{\text{fare, £5K<inc}< £10K}$	-2.0727	-1.91
β_{tt}	-0.0377	-8.44	$\beta_{\text{fare, £10K<inc}< £15K}$	-1.7193	-2.46
β_{crowding}	-1.7759	-7.44	$\beta_{\text{fare, £15K<inc}< £20K}$	-0.6150	-1.16
$\beta_{\text{rate of delays}}$	-1.8937	-8.04	$\beta_{\text{fare, £20K<inc}< £30K}$	-2.1777	-7.10
$\beta_{\text{av. delay}}$	-0.0344	-4.75	$\beta_{\text{fare, £30K<inc}< £40K}$	-1.0986	-2.02
$\beta_{\text{delay sms}}$	0.2774	4.67	$\beta_{\text{fare, £40K<inc}< £50K}$	-0.7794	-1.25
			$\beta_{\text{fare, £50K<inc}< £75K}$	-0.4418	-2.15
			$\beta_{\text{fare, inc}> £75K}$	-0.7527	-1.96
			$\beta_{\text{fare, inc unknown}}$	-0.6370	-3.25
			$\beta_{\text{fare, inc refused}}$	-1.2661	-2.66

Table 3: Estimation results for MMNL model for first case study

log-likelihood -2,992.65

	est.	rob. t-rat		est.	rob. t-rat
δ_1	0.792	9.95			
δ_2	0.333	4.49			
$\mu\left(ln\left(-\beta_{\text{tt}}\right)\right)$	-2.637	-23.17	$\sigma\left(ln\left(-\beta_{\text{tt}}\right)\right)$	-0.823	-9.61
$\mu\left(ln\left(-\beta_{\text{fare}}\right)\right)$	3.845	3.34	$\sigma\left(ln\left(-\beta_{\text{fare}}\right)\right)$	1.598	18.98
$\mu\left(ln\left(-\beta_{\text{fare no inc.}}\right)\right)$	0.899	5.66	$\sigma\left(ln\left(-\beta_{\text{fare no inc.}}\right)\right)$	1.884	9.84
$\mu\left(ln\left(-\beta_{\text{crowding}}\right)\right)$	0.756	3.73	$\sigma\left(ln\left(-\beta_{\text{crowding}}\right)\right)$	-1.626	-15.41
$\mu\left(ln\left(-\beta_{\text{rate of delays}}\right)\right)$	1.044	7.21	$\sigma\left(ln\left(-\beta_{\text{rate of delays}}\right)\right)$	1.098	12.07
$\mu\left(ln\left(-\beta_{\text{av. delay}}\right)\right)$	-3.316	-9.65	$\sigma\left(ln\left(-\beta_{\text{av. delay}}\right)\right)$	-1.371	-9.54
$\mu\left(ln\left(\beta_{\text{delay sms}}\right)\right)$	-1.143	-5.47	$\sigma\left(ln\left(\beta_{\text{delay sms}}\right)\right)$	1.221	14.45
λ_{inc}	-0.253	-2.27			

measurement model and is thus not directly comparable to the earlier models. The model fit relating to the choice model component within the hybrid model is indistinguishable from that of the simple MMNL model given possible simulation error. In fact, lower likelihood for the choice model component could be expected as, in the hybrid model, the optimisation is not focussed on the choice model alone - in our case, the deterioration is negligible. Additionally,

Table 4: Estimation results for choice model component within the latent variable model for first case study

			overall log-likelihood:	-3,583.58	
			log-likelihood for choice model component:	-2,992.89	
	est.	rob. t-rat			
δ_1	0.787	9.77			
δ_2	0.335	4.53		est.	rob. t-rat
$\mu\left(ln\left(-\beta_{\text{tt}}\right)\right)$	-2.652	-24.31	$\sigma\left(ln\left(-\beta_{\text{tt}}\right)\right)$	-0.839	-8.52
$\mu\left(ln\left(-\beta_{\text{fare}}\right)\right)$	1.417	14.77	$\sigma\left(ln\left(-\beta_{\text{fare}}\right)\right)$	1.536	12.29
$\mu\left(ln\left(-\beta_{\text{crowding}}\right)\right)$	0.673	2.52	$\sigma\left(ln\left(-\beta_{\text{crowding}}\right)\right)$	-1.860	-6.35
$\mu\left(ln\left(-\beta_{\text{rate of delays}}\right)\right)$	1.038	6.40	$\sigma\left(ln\left(-\beta_{\text{rate of delays}}\right)\right)$	1.027	10.13
$\mu\left(ln\left(-\beta_{\text{av. delay}}\right)\right)$	-3.222	-10.76	$\sigma\left(ln\left(-\beta_{\text{av. delay}}\right)\right)$	-1.294	-9.86
$\mu\left(ln\left(\beta_{\text{delay sms}}\right)\right)$	-1.056	-5.53	$\sigma\left(ln\left(\beta_{\text{delay sms}}\right)\right)$	1.124	11.55
λ_{inc}	-0.201	-2.85			

the simple MMNL model uses an additional fare coefficient for respondents with missing income; if anything, the fact that the log-likelihood is so similar is thus an endorsement of the hybrid model.

Table 4 contains the estimation results for the choice model component. The estimates are in line with the findings for the MMNL model, albeit that the estimated income elasticity is slightly lower still. The difference is that this model no longer relies on stated income, but instead uses a latent variable. As shown in Table 5, the structural equation for this latent variable has a number of statistically significant socio-demographic effects, with higher latent income for rail users, older respondents, respondents with university degrees and respondents who have a car available, while it is lower for female respondents and respondents aged under 35. The estimates for these socio-demographic effects are informed by the stated income for those respondents where it is available, and by the observed choices for all respondents.

In the measurement model, we explain stated income (where available) through an ordered logit model, with eight estimated thresholds, and where ζ gives the impact of the latent variable in this ordered logit model, with the positive estimate showing that as latent income increases, so does the stated income. This means that the effects are consistent in the two model components. A higher value for the latent income variable leads to lower cost sensitivity in the choice model (negative income elasticity), while, in the measurement model, it also leads to a higher probability for stated income to be in the higher categories (ζ). The results are overall remarkably consistent with those from MMNL, pos-

Table 5: Estimation results for structural equation and measurement model component within latent variable model for first case study

	est.	rob. t-rat
γ_{train}	0.6624	3.71
γ_{female}	-0.5056	-3.09
$\gamma_{undegraduate}$	0.8559	4.42
$\gamma_{postgrad.}$	1.2476	6.29
$\gamma_{under35}$	-0.3598	-2.45
γ_{over55}	0.1915	1.39
$\gamma_{caravailable}$	0.7785	5.35
ζ	0.6152	2.39
threshold 1	-2.0353	-5.56
threshold 2	-0.4994	-1.57
threshold 3	0.4816	1.38
threshold 4	1.6090	3.58
threshold 5	3.8471	4.43
threshold 6	5.7575	4.16
threshold 7	8.9056	3.94
threshold 8	15.821	3.48

sibly as a result of the low rate of missing data and narrow income classes. On the flip side, this is a clear indication that the proposed model *works*.

To conclude the discussion of the first case study, Table 6 shows the implied sample population level distributions for the willingness-to-pay (WTP) for improvements in the five non-cost attributes, where we focus on the 30^{th} and 70^{th} percentiles, the median and the mean. For the simple MNL model, there is no heterogeneity, while, in the MNL model with an income interaction, we see that the WTP at the 70^{th} percentile is 2.66 times as high as that at the 30^{th} percentile point. In the MMNL model, the heterogeneity is increased substantially, while we also see increases in the mean levels, arguably to more sensible values. Except for crowding, the mean levels as well as the amount of heterogeneity is lower in the hybrid model, but still closer to the MMNL findings than the MNL results. While simple random heterogeneity not related to income dominates the WTP patterns, the findings still suggest a non-trivial impact by the latent income variable.

4 Second case study

Our second case study makes use of revealed preference data from a survey for car ownership in Japan. The data come from the Japanese General Social

Table 6: Implied sample population level willingness to pay distributions for first case study

	tt (/hr)			
	30^{th} percentile	median	mean	70^{th} percentile
MNL	2.16	2.16	2.16	2.16
MNL with income	1.09	1.79	2.12	2.90
MMNL	0.45	1.18	6.92	3.11
Hybrid	0.48	1.23	6.08	3.15

	crowding (per 1/10 train)			
	30^{th} percentile	median	mean	70^{th} percentile
MNL	0.17	0.17	0.17	0.17
MNL with income	0.09	0.14	0.17	0.23
MMNL	0.02	0.06	0.88	0.20
Hybrid	0.02	0.06	1.08	0.21

	rate of delays (per 1/10 train)			
	30^{th} percentile	median	mean	70^{th} percentile
MNL	0.18	0.18	0.18	0.18
MNL with income	0.09	0.15	0.18	0.24
MMNL	0.03	0.08	0.59	0.22
Hybrid	0.03	0.08	0.48	0.22

	av delay (/hr)			
	30^{th} percentile	median	mean	70^{th} percentile
MNL	2.04	2.04	2.04	2.04
MNL with income	1.00	1.63	1.93	2.65
MMNL	0.20	0.60	6.32	1.85
Hybrid	0.24	0.70	5.57	2.03

	delay sms provision			
	30^{th} percentile	median	mean	70^{th} percentile
MNL	0.27	0.27	0.27	0.27
MNL with income	0.13	0.22	0.26	0.36
MMNL	0.03	0.09	0.77	0.26
Hybrid	0.04	0.10	0.66	0.28

Survey 2005 (JGSS-2005) collected in Japan in 2005, which is a part of the JGSS series started in 2000. The survey area covers all of Japan, and the sample includes respondents aged between 20 and 89. The interview collects information including gender and age of all household members, car ownership and income (19 categories) on a household level, occupation and education for the respondent and his/her spouse, and various others. For the present study, we made use of a sample of $1,668$ respondents who provided information on car

Table 7: Estimation results for a binary probit model without income for second case study

log-likelihood: -474.65

	est.	rob. t-rat.
number of males (-17yrs) in household	0.8919	4.52
number of males (18-yrs) in household	0.9112	9.01
number of females (-17yrs) in household	0.5753	4.69
number of females (18-yrs) in household	0.5959	7.25
Tokyo's 23 wards	-1.1250	-6.17
Yokohama/Kawasaki cities	-0.7100	-2.33
Osaka city	-1.2708	-3.09
threshold	0.7230	4.78

ownership. Out of this sample, income was missing for 614 respondents, i.e. a very substantial 36.8% of the sample.

We first estimated a simple binary probit model, with results summarised in Table 7. The dependent variable is a binary indicator for whether a household owns a car or not, meaning that we estimate a single threshold in the model. We use four socio-demographic effects, showing that increases in household size lead to increased probability of owning cars, where this effect is stronger for the number of male household members. The number of household members aged 18 years old or older, who can obtain a driving licence, was expected to have a larger impact than the number of younger members, but the estimated effects were similar. Three variables relating to urban areas are included, showing that the probability of owning cars is smaller in large cities, where good accessibility to public transport is provided.

We contrast this model with a specification which includes a continuous income effect in Table 8, using the category mid-points for households with reported income, and a separate effect for households with missing income. This leads to a highly significant improvement in model fit by 28.06 units of log-likelihood at the cost of two additional parameters. The results show that as income increases, so does the probability of car ownership, where a constant is estimated for respondents with no reported income.

We finally turn our attention to the latent model, with results summarised in Tables 9 and 10. In the measurement model, we use an ordered probit model with 18 thresholds, where we note that, as the latent income increases, so does the probability of higher stated income, given the positive estimate

Table 8: Estimation results for a binary probit with continuous income for second case study

log-likelihood: -446.59

	est.	rob. t-rat.
number of males (-17yrs) in household	0.8536	4.56
number of males (18-yrs) in household	0.7873	7.87
number of females (-17yrs) in household	0.4878	4.03
number of females (18-yrs) in household	0.4906	5.77
Tokyo's 23 wards	-1.2031	-6.46
Yokohama/Kawasaki cities	-0.8727	-2.82
Osaka city	-1.4387	-3.66
threshold	1.0370	6.60
annual household income (JPY 10 million) if reported	1.2910	4.54
income missing	0.7007	4.63

for ζ. Unlike in the previous case study where the latent income was used in interaction with the fare coefficient, the latent income in this case study enters directly into the utility function, where in the present exploratory work, we use a linear component of the form $\tau\alpha_{nk}$. We note the positive impact that the latent income variable has on car ownership as evidenced by the positive estimate for τ.

In terms of model fit, the share of the log-likelihood relating to the binary probit component is -438.67 compared to -446.59 for the simple binary probit model with continuous income in Table 8. In the absence of a formal statistical test, this still suggests that the choice model component within the hybrid model gives a better representation of the car ownership than a model optimised solely for that part of the data, where the latter also made use of an additional parameter for respondents with missing income. This provides a further indication of the potential benefits of the latent approach, where some of the gains are of course a result of explaining latent income through a wide range of socio-demographic variables.

Those parameters shared with the binary probit model remain broadly comparable, but the income effect now obtains a higher level of statistical significance, no doubt helped in part by the fact that it is now used across all respondents, including those whose income was not reported (i.e. where a separate effect was used in Table 8).

Turning to the structural equation for the latent variable, the latent income

Table 9: Estimation results for the binary probit component of the latent variable model for second case study

overall log-likelihood:	-2,974.56
log-likelihood for binary probit component:	-438.67

	est.	rob. t-rat.
number of males (-17yrs) in household	0.7727	4.33
number of males (18-yrs) in household	0.7631	8.03
number of females (-17yrs) in household	0.3442	2.81
number of females (18-yrs) in household	0.5299	5.80
Tokyo's 23 wards	-1.3389	-6.64
Yokohama/Kawasaki cities	-0.8456	-2.58
Osaka city	-1.3992	-3.57
threshold	1.1316	6.27
τ	0.3908	5.66

is explained by characteristics of the respondent and his/her spouse[3]. The respondent is either male without wife, female without husband, or male or female with his/her spouse, so only two constants are included. Note that unmarried male and female respondents also are termed husband and wife respectively in this case study. The effects of working, alone and in interaction with age, on latent income are examined; the former is statistically significant for only wives, while the latter is statistically significant only for husbands. Both high school and university education lead to higher latent income, and the effect of the latter is larger (base is no high school education). Working hours have an impact on latent income especially for wives. Employment status is statistically significant only for husbands, with the highest effect for executives, followed by department head and section head (the base is lower than the section head level). Latent income is higher for husbands employed in large companies or government agencies compared to those in smaller companies. Husbands employed in finance/insurance industries have higher latent income compared to other industries. In summary, a husband's pay depends on the type of employment and experience, while a wife's pay depends more on the hours worked, with less variance in hourly rates (strong per-hour effect).

In this data, unlike the first set, it is not possible to compute measures of

[3]Since detailed socio-demographic information is available for only the respondent and his/her spouse, the latent income is explained by their characteristics. On the other hand, car ownership is explained by characteristics of all household members. Investigation of which variables should appear in each of the submodels is a task for further research.

Table 10: Estimation results for the structural equation for latent income and for the measurement model for second case study

	est.	rob. t-rat.
husband in household	0.5813	2.61
wife in household	0.0925	0.42
wife if working	0.3917	1.75
age of husband if he works	0.0077	2.22
high school education for husband	0.2982	1.73
high school education for wife	0.4758	2.81
university education for husband	0.6800	3.32
university education for wife	0.7932	3.84
working hours per week for husband	0.0072	1.79
working hours per week for wife	0.0181	2.91
husband working as executive	1.8531	5.97
husband working as department head	1.1486	4.61
husband working as section head	0.7442	3.60
husband working as large company employee	0.6870	4.39
husband working as government employee	1.0061	4.69
husband working as in financial institutions/insurance	1.0446	3.05

	est.	rob. t-rat.		est.	rob. t-rat.
threshold 1	-1.5740	-6.81	threshold 10	2.5797	9.47
threshold 2	-0.9451	-4.90	threshold 11	2.9292	10.07
threshold 3	-0.5494	-3.02	threshold 12	3.2370	10.51
threshold 4	-0.2024	-1.13	threshold 13	3.7503	11.07
threshold 5	-0.0184	-0.10	threshold 14	4.1494	11.40
threshold 6	0.5637	2.97	threshold 15	4.5139	11.61
threshold 7	1.2094	5.76	threshold 16	4.8804	11.64
threshold 8	1.7690	7.59	threshold 17	5.1553	11.65
threshold 9	2.2258	8.72	threshold 18	5.4996	12.15
ζ	0.7413	5.90			

willingness to pay because of the absence of attributes relating to the alternatives. However, it is possible to calculate an income elasticity of car ownership to allow quantitative comparisons to be made between the outputs of the alternative models. In Table 11, results are shown for the binary probit model, where it is possible to estimate this statistic only for the respondents who gave their income, and for the hybrid model, separately for the two groups of respondents and for the whole data set. It can be seen that the results for those respondents

Table 11: Income elasticity of owning one or more cars

	Elasticity
Binary probit (respondents giving income)	0.093
Hybrid (respondents giving income)	0.098
Hybrid (respondents not giving income)	0.065
Hybrid (all respondents)	0.085

reporting income match quite closely between the binary probit and the hybrid model, which is reassuring. However, the binary model is not able to provide an elasticity for the non-reporters, which we know from the hybrid model to be substantially lower than that for reporters, by around a third. As a result, the overall elasticity for the entire sample in the hybrid model is corrected downwards from the prediction for those respondents who report income, This gives a clear indication of the advantage of the hybrid structure. Finally, overall, the elasticity might appear low, but it must be remembered that this is the elasticity for having one or more cars, not total car ownership, in a market where 86% of the households have cars.

5 Conclusions

This paper has looked at the possibility of treating income as a latent variable in a choice model context. The motivation for such an approach is that stated income is affected by a number of key issues, in the form of measurement error (e.g. due to being measured in categories), missing observations, correlation with other unobserved factors, and bias introduced by the respondent in the form of under or overstated income. In our specification, the latent income variable is then used inside the choice model as well as being used as an explanator in a measurement model used to explain stated income. The latent income variable has a deterministic as well as a random component, where the former explains latent income as a function of other respondent characteristics.

Unlike a method relying directly on stated income, this approach has the advantage of making provision for error in the respondent's reported income. By treating stated income as a dependent rather than explanatory variable, it also avoids issues with endogeneity bias. In comparison with using imputation for missing income information, the method has the advantage that the latent income variable is informed not just by the relationship between observed income (i.e. for those respondents who provide it) and other socio-demographic characteristics, but also by the choices made by all respondents.

The method is directly applicable for forecasting. This is in contrast with using stated income, where an analyst needs to either discard a large share of their data, i.e. respondents with missing income, or use imputed values for all. Indeed, the approach of estimating separate cost sensitivites for respondents with missing income does not easily carry over into forecasting when the forecast population is different from the estimation sample.

In summary, it appears from these tests that treating income as a latent variable is a viable and useful method for analysing this type of data. The results for the choice model are comparable with those for other approaches, possibly better when there is a substantial amount of data with missing income. Possibilities of imputation bias are reduced, while the method is simpler to use than some alternatives. Additionally, the benefits over other methods in forecasting are obvious.

The work presented in this paper offers a number of opportunities for further development. First, substantial work is still required on the functional form for the measurement model, the structural equation of the latent income variables, as well as the role of them in the choice model. Second, while, unlike traditional approaches, no specific treatment is needed for respondents with missing income in the choice model, the current assumption of treating them in the same manner as other respondents might need to be revisited, albeit that they already do not contribute to measurement model, only to choice. Third, the rationale for using specific variables in the choice model and structural model needs to be improved. These are important directions for future work, but the present paper has given a clear indication of the applicability of the method.

Acknowledgements

The Japanese General Social Surveys (JGSS) are designed and carried out at the Institute of Regional Studies at Osaka University of Commerce in collaboration with the Institute of Social Science at the University of Tokyo under the direction of Ichiro TANIOKA, Michio NITTA, Noriko IWAI and Tokio YASUDA. The project is financially assisted by Gakujutsu Frontier Grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology for 1999-2008 academic years, and the datasets are compiled and distributed by SSJ Data Archive, Information Center for Social Science Research on Japan, Institute of Social Science, the University of Tokyo. The second author acknowledges the financial support from a Grant-in-Aid for Scientific Research (Grant Nos. 22730334 and 25380564) from the Japan Society for the Promotion of Science.

References

Ashok, K., Dillon, W. R., Yuan, S., 2002. Extending discrete choice models to incorporate attitudinal and other latent variables. Journal of Marketing Research, 31–46.

Ben-Akiva, M., Walker, J., Bernardino, A., Gopinath, D., Morikawa, T., Polydoropoulou, A., 2002a. Integration of choice and latent variable models. In: Mahmassani, H. (Ed.), In Perpetual motion: Travel behaviour research opportunities and application challenges. Pergamon, Ch. 13, pp. 431–470.

Ben-Akiva, M., Walker, J., McFadden, D., Gärling, T., Gopinath, D., Bolduc, D., Börsch-Supan, A., Delquié, P., Larichev, O., Morikawa, T., Polydoropoulou, A., Rao, V., 1999. Extended framework for modeling choice behavior. Marketing Letters 10 (3), 187–203.

Ben-Akiva, M., Walker, J., McFadden, D., Train, K., Bhat, C. R., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D. S., Daly, A., de Palma, A., Gopinath, D., Karlstrom, A., Munizaga, M. A., 2002b. Hybrid choice models: Progress and challenges. Marketing Letters 13 (3), 163–175.

Bhat, C., 1994. Imputing a continuous income variable from grouped and missing income observations. Economics Letters 46, 311–319.

Bierlaire, M., 2003. BIOGEME: a free package for the estimation of discrete choice models. Proceedings of the 3^{rd} Swiss Transport Research Conference, Monte Verità, Ascona.

Bierlaire, M., 2005. An introduction to BIOGEME Version 1.4. biogeme.epfl.ch.

Bolduc, D., Ben-Akiva, M., Walker, J., Michaud, A., 2005. Hybrid choice models with logit kernel: Applicability to large scale models. In: Lee-Gosselin, M., Doherty, S. (Eds.), Integrated Land-Use and Transportation Models: Behavioural Foundations. Elsevier, Oxford, pp. 275–302.

Brey, R., Walker, J., 2011. Latent temporal preferences: An application to airline travel. Transportation Research Part A 45 (9), 880–895.

Brownstone, D., Steimetz, S., 2005. Estimating commuters vot with noisy data. Transportation Research Part B 39 (10), 865–889.

Choudhury, C., Ben-Akiva, M., Abou-Zeid, M., 2010. Dynamic latent plan models. Journal of Choice Modelling 3 (2), 50–70.

Daly, A., Zachary, S., 1977. The Effect of Free Public Transport on The Journey to Work. Transport and Road Research Laboratory Report SR388.

Hess, S., Stathopoulos, A., Daly, A. J., 2012. Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. Transportation 39 (3), 565–591.

Walker, J., Li, J., Srinivasan, S., Bolduc, D., 2010. Travel demand models in the developing world: correcting for measurement errors. Transportation Letters 2, 231–243.