The Journal of Choice Modelling 12 (2014) 47–57

Contents lists available at ScienceDirect

The Journal of Choice Modelling

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

Contrasting imputation with a latent variable approach to dealing with missing income in choice models

Nobuhiro Sanko [a,b,*], Stephane Hess [b], Jeffrey Dumont [b,c], Andrew Daly [b,d]

[a] *Graduate School of Business Administration, Kobe University, Japan*
[b] *Institute for Transport Studies, University of Leeds, United Kingdom*
[c] *RSG, United States*
[d] *RAND Europe, United Kingdom*

ARTICLE INFO

Keywords:
Latent variables
Missing income
Discrete choice
Random heterogeneity

ABSTRACT

Income is a key variable in many choice models. It is also one of the most salient examples of a variable affected by data problems. Issues with income arise as measurement errors in categorically captured income, correlation between stated income and unobserved variables, systematic over- or under-statement of income and missing income values for those who refuse to answer or do not know their (household) income. A common approach for dealing especially with missing income is to use imputation based on the relationship among those who report income between their stated income for reporters and their socio-demographic characteristics. A number of authors have also recently put forward a latent variable treatment of the issue, which has theoretical advantages over imputation, not least by drawing not just on data on stated income for reporters, but also choice behaviour of all respondents. We contrast this approach empirically with imputation as well as simpler approaches in two case studies, one with stated preference data and one with revealed preference data. Our findings suggest that, at least with the data at hand, the latent variable approach produces similar results to imputation, possibly an indication of non-reporters of income having similar income distributions from those who report it. But in other data sets the efficiency advantage over imputation could help in revealing issues in the complete and accurate reporting of income.

1. Introduction

Income is arguably the most important socio-demographic variable in terms of explaining deterministic heterogeneity across respondents in choice models, notably in terms of explaining variations in cost sensitivity. It is however also one of the most difficult measures to capture accurately. Surveys often suffer from high rates of non-reporting for income, primarily due to respondents refusing to provide this information but also due to an often non-trivial share of respondents indicating that their actual total income is not known to them. Rates of non-reporting tend to be lower when capturing income as a categorical (rather than continuous) variable, and especially when using broader intervals, but this inevitably leads to measurement error. In addition to missing income information for some respondents, there are also potential issues in terms of correlation between stated income and other unobserved factors, as well as systematic respondent caused error, for example in the form of under- or over-reporting.

[*] Corresponding author. Tel. +81 78 803 6987.
E-mail addresses: sanko@kobe-u.ac.jp (N. Sanko), s.hess@its.leeds.ac.uk (S. Hess), jeff.dumont@rsginc.com (J. Dumont), daly@rand.org (A. Daly).

http://dx.doi.org/10.1016/j.jocm.2014.10.001
1755-5345/© Elsevier Ltd. All rights reserved.

The issues of measurement error and missing variables seem rarely to be addressed in practical choice modelling (with Walker et al., 2010 being one exception), although it is clear that such issues are likely to lead to higher error in a choice model and sometimes cause bias in parameter estimates. Hausman (2001) draws attention to this issue, for both linear and non-linear models and for both right-side and left-side variables. He also suggests solutions based on previous work, though the approach of this paper is closer to that of Walker et al. (2010) than that of Hausman (2001). Implicitly, measurement error is often ignored in practice, with an assumption that its effects are captured in the error term of the model. This can be more serious if, as is probable, measurement error or non-reporting is 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. 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. 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, potentially leading to biased results. Studies that analyse measurement errors include Walker et al. (2010), who introduced latent level-of-service variables to account for reported level-of-service variables with measurement error. Correlation with other unobserved factors can cause bias, and despite important work by for example Petrin and Train (2010) and Guevara and Ben-Akiva (2010), many studies still ignore the potential risk of such correlation, especially when it concerns explanatory variables in revealed preference (RP) data.

With income, the main focus in practical work has been on the treatment of missing income. A still all too common approach is to remove affected respondents from the data, which obviously leads to an undesirable reduction in sample size, can make the resulting dataset less representative of the real population, and cause endogeneity issues because of self-selection. These factors all have implications both in terms of computing willingness-to-pay (WTP) measures and forecasting. A crude approach is to place non-reporters at the sample-level mean income for reporters, but this assumption may not be justified and it is safer to estimate a separate cost coefficient for non-reporters. This allows a model to show whether non-reporters have higher (or lower) than average cost sensitivity, implying lower (or higher) than average income. Although this approach allows non-reporters to be incorporated in sample-level calculations of WTP measures, problems arise in forecasting, primarily as it becomes difficult to formulate the impact of income changes at the sample-level, given the special treatment for non-reporters.

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 (e.g. Jiang and Morikawa, 2007). 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. 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, previously treated analytically by Daly and Zachary (1977) or by ‘multiple imputation’ by Rubin (1987) and Brownstone and Steimetz (2005).

Imputation is also only informed by the observed values for the missing variable for other respondents, and not for example by the observed choice behaviour of respondents with missing data. This relates to the point about the actual income for non-reporters potentially being different from that of reporters who have otherwise similar characteristics. Additionally, with imputation, analysts often still treat the income for reporters as error free measures, when in reality, this will not be appropriate, especially when income is captured as a categorical variable.

In recent years, a number of applications have put forward the treatment of income as a latent variable, notably in the examples of the BIOGEME software (Bierlaire, 2003, 2005) and in Bolduc and Alvarez-Daziano (2010). This leads to the use of a hybrid model framework, an approach that is 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 key concept is that the variable of interest is considered as being unobserved, with only indicators thereof being captured in the data. A structural equation is employed to characterise the latent variable, and this is used to explain both the values of the indicators and the role of the latent component in the choice model. 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), or the treatment of level-of-service (Walker et al., 2010), preferred arrival time (Brey and Walker, 2011) and budgets (Dumont et al., 2013) as latent variables.

The use of hybrid models in the present context relies on formulating a single latent variable that represents a continuous income measure, which is a function of a number of other socio-demographic characteristics as well as a random disturbance. This latent income variable is then used to explain the stated income for those respondents who reported it and is also used as a measure of income inside the utility functions of the choice model, for example to explain variations in cost sensitivity.

The hybrid approach has a number of potential key advantages over the alternative methods discussed above. In common with using imputation, the models are directly applicable for computing WTP distributions for the entire sample and also for using all respondents in forecasting, given that an income variable is now available for every decision maker. In contrast with imputation, the hybrid model however no longer treats stated income as an error-free measure of real income, potentially giving it an advantage in terms of accommodating measurement error. Furthermore, as stated income is no