Estimation of Travel Demand Models with Grouped and Missing Income Data

CHANDRA BHAT

A method to impute a continuous value for household income from grouped and missing income data for use as an explanatory variable in travel demand estimation was developed. Many data sets collect income in a discrete number of categories or in grouped form to simplify the respondent's task and to encourage a response. In spite of such grouped data collection, many respondents refuse to provide information on income, leading to missing income values. The issue of constructing a continuous measure of income from grouped and missing income data that, when used in travel demand models as an explanatory variable, enables consistent estimation of the model parameters is addressed.

Household income is an important sociodemographic explanatory variable in travel demand models such as car ownership models (*1*), trip generation models (*2*), and mode choice models (*3*). In almost all transportation data sets and in many other data sets (*4*) household income, an inherently continuous variable, is measured in a discrete number of categories or intervals; that is, it is measured in grouped form (e.g., between $15,000 and $30,000). The income question is also notorious for its high nonresponse rates, leading to missing income observations in most data sets.

Income is measured in grouped form for two related reasons. First, such a measuring scale provides a greater degree of protection of confidentiality compared with a continuous measure (the degree of protection being a function of the size of income intervals), thereby increasing response rates (*5*). Second, it renders the sensitive income question relatively innocuous during survey administration. Questions that seek a continuous measure on income can offend respondents, particularly in a telephone survey or in a personal interview survey in which respondents are put "on the spot."

Although income is measured in grouped form, it is the continuous measure of income (or some function of this continuous measure) that frequently appears as an explanatory variable in travel demand models. It is important that this continuous measure be a reliable measure of the true income value to enable the development of an accurate and reliable relationship between travel demand variables and their explanatory variables and thus facilitate good prediction of travel demand variables [the research by Hamburg et al. (*6*) indicates that the estimates in a travel demand model are highly sensitive to the accuracies of sociodemographic input variables and emphasizes the need for accurate measures of the input variables]. This paper proposes a method for constructing such a continuous measure of income for all observations in a cross-sectional data set with grouped and missing income data.

The next section of this paper discusses the motivation for developing methods to explicitly accommodate the grouped and missing nature of income data in travel demand modeling. The subsequent section presents the need to develop a model relating household income and factors affecting household income to impute a continuous income measure from grouped and missing income data for use as an exogenous variable in travel demand models. The following section advances an econometric framework used to impute a continuous income measure through the development of a model relating income to variables influencing income. Empirical results obtained by using a Dutch data set are then presented. The final section provides a summary of the research and highlights important findings.

Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, Mass. 01003.

MOTIVATION FOR TREATMENT OF GROUPED AND MISSING INCOME DATA

The motivation for the treatment of grouped and missing income data originates from the need to develop a consistent relationship between travel demand variables and their explanatory variables (including income). The dependent variable in the demand model may be an observed continuous variable such as trip generation or a latent continuous variable that is a reflection of an observed discrete choice decision such as utilities in the case of a mode choice decision or car ownership propensity in the case of an ordered car ownership model. Unfortunately current procedures for constructing a continuous measure from grouped data and commonly used techniques for handling missing income data do not enable consistent estimation of travel demand models. This inconsistency in commonly used demand estimation procedures without and with missing income data is discussed below.

Commonly Used Estimation Procedures

Grouped Without Missing Income Data

Commonly used estimation procedures construct a continuous value of income from grouped data by assigning the midpoint of the income threshold bounds that determine each category to each observation in that category. If the threshold bounds for income category j are a_{j-1} (the lower bound) and a_j (the upper bound), then a continuous income value ζ is constructed for all observations in category j as

$$\zeta_i|(\text{income category} = j) = \frac{a_{j-1} + a_j}{2} \quad (1)$$

In the case of the two categories at either end of the income spectrum, an arbitrary truncation point is used as the representative value.

This midpoint method of constructing a continuous measure from grouped data has serious limitations. Consider an underlying linear regression between a demand variable y_i and the actual (but unobservable) income variable I_i^* as follows [the following presentation is based on Hsiao (*7*) and is confined to the case when the dependent demand variable is an observed continuous variable for ease in presentation]:

$$y_i = \alpha + \beta I_i^* + u_i \tag{2}$$

where

i = index for observations,
α and β = parameters to be estimated, and
u_i = an error term.

Assume the standard regression conditions that u_i is an independent and identically distributed (iid) random error term with a mean of zero and I_i^* is uncorrelated with the error term. If the actual income value I_i^* for an observation is replaced by the midpoint of the corresponding income category, the regression may be rewritten as

$$y_i = \alpha + \beta\zeta_i + v_i \tag{3}$$

where $v_i = u_i + \beta(I_i^* - \zeta_i)$. In this case when the midpoint values are used, the coefficient of β is given by (using Equation 2)

$$\hat{\beta}_{mid} = \frac{\sum_i (y_i - \bar{y})(\zeta_i - \bar{\zeta})}{\sum_i (\zeta_i - \bar{\zeta})^2} = \beta \frac{\sum_i (I_i^* - \bar{I}^*)(\zeta_i - \bar{\zeta})}{\sum_i (\zeta_i - \bar{\zeta})^2} \tag{4}$$

To simplify this expression write the actual (but unobserved) continuous income I_i^* for an observation i falling in the grouped income category j as the sum of three components: the midpoint of the category j, ζ_i, as computed in Equation 1; an error term τ_i representing the difference between the expected value of I_i^* given that it falls in category j (or the expected value of the marginal distribution of the continuous income variable between the threshold bounds of category j) and the midpoint of category j; and a random error term, w_i, representing the difference between the actual continuous income I_i^* and the expected value of I_i^* given that it falls in category j. That is,

$$I_i^* = \zeta_i + \tau_i + w_i \tag{5}$$

where $\tau_i = E[I_i^*|\text{cat.}j] - \zeta_i$ and $w_i = I_i^* - E[I_i^*|\text{cat.}j]$. By using Equation 5 one can write $I_i^* - \bar{I}^* = (\zeta_i - \bar{\zeta}) + (\tau_i - \bar{\tau})$. By substituting this expression into Equation 4 one can rewrite the least-squares estimate of β by the midpoint method as

$$\underset{N\to\infty}{\text{Plim}}\ \hat{\beta}_{mid} = \beta + \beta \frac{\text{Cov}(\tau_i, \zeta_i)}{\text{Var}(\zeta_i)} \tag{6}$$

Thus the parameter estimate on income obtained by the midpoint method converges to the actual value of β in the travel demand model if and only if $\text{Cov}(\tau_i, \zeta_i)$ converges to zero. However this will generally not be the case. The magnitude and direction of $\text{Cov}(\tau_i, \zeta_i)$ depend on the shape and distribution of the actual (but unobserved) income variable. Earlier studies (*8,9*) have indicated that a log-normal form is theoretically and empirically appropriate for the income distribution. $\text{Cov}(\tau_i, \zeta_i)$ is, in general, not equal to zero for a log-normal distribution. No general result regarding the direction and magnitude of $\text{Cov}(\tau_i, \zeta_i)$ (and therefore the direction and magnitude of the bias of the midpoint method) can be established for the log-normal distribution. A more definitive result can be established if it is assumed that I_i^* in Equation 2 represents the logarithm transformation of actual income. In this case I_i^* is normally distributed (since actual income is log-normally distributed). Assuming small tail distributions, τ_i decreases from a positive value for the lower income categories (the expected value of the normal distribution between the threshold bounds of category j is greater than the midpoint) to a negative value for the higher income categories (the expected value of the distribution between the threshold bounds of category j is lower than the midpoint) as indicated by Haitovsky (*10*). On the other hand the midpoint of income categories increases as one proceeds from lower to higher categories. Thus the covariance term, $\text{Cov}(\tau_i, \zeta_i)$, is negative and the midpoint estimate $\hat{\beta}_{mid}$ in Equation 6 underestimates β.

The midpoint method leads to inconsistent parameter estimates (a parameter estimate $\hat{\beta}$ is said to be a consistent estimator of the true β if, as the sample size gets infinitely large, the probability that $|\beta - \hat{\beta}|$ will be less than any arbitrary small positive number approaches 1) in the travel demand model because τ_i is not equal to zero. However if a consistent imputed estimate of income (that is, a consistent estimate of the expected value of I_i^* given that it falls in category j) is used instead of the midpoints, τ_i is zero and one obtains consistent parameters in the travel demand model (the reader will observe that as the number of income categories increases, or more appropriately as the size of the income interval within each income category decreases, τ_i becomes closer to zero in the midpoint method and the inconsistency resulting from use of the midpoint method is reduced).

The results regarding the inconsistency of the midpoint method are generalizable to the case of many explanatory variables in the travel demand model. Specifically use of the midpoint income estimate as an explanatory variable leads to inconsistent parameter estimates on all of the explanatory variables in the model, not just the income variable (*7*).

Grouped with Missing Income Data

The discussion above assumed that there are no missing income observations. Now consider the limitations of commonly used methods when missing income data are present. Current methods adopt one of two strategies to estimate travel demand models from grouped and missing income data. The first strategy is to assign the midpoint of income categories for observations with observed (grouped) income values and to assign the average value of the midpoint estimates of the observed income observations to the missing income observations. As discussed earlier the midpoint method does not provide consistent estimates of the travel demand model. Also this assignment of the average of observed income observations to missing income observations assumes that the average income of respondent households (i.e., households that report income) is identical to that of nonrespondent households (i.e., households that do not report income). This may not be true because of systematic variations in observed and unobserved characteristics affecting income earnings between members of respondent and nonrespondent households (*11*). Observed characteristics

may include the education levels of the members of the household, whereas unobserved characteristics may include sensitivity to privacy and fear of governmental or other uses of the data. If such systematic variations are present between members of respondent and nonrespondent households, assigning the average income of respondent households to nonrespondent households is inappropriate and will further contribute to inconsistency in the parameter estimates of the demand model.

The second strategy for estimating travel demand models from grouped and missing income data is to assign the midpoint of income category thresholds for the observed (grouped) income data and to drop all missing income observations. It was already shown that the midpoint method provides inconsistent travel demand parameters. In addition another dimension of inconsistency arises when all missing income observations are dropped. If systematic variations in income level are present between respondent and nonrespondent households, then the relationship between independent variables and the travel demand variable for nonrespondents may be different from that for respondents. Thus the travel demand relationship obtained by dropping all nonrespondent households will not be a representative relationship for the entire population. This second strategy of dropping missing income observations also results in a loss of observations, resulting in inefficient estimation.

It is clear from the discussion above that commonly used procedures for dealing with grouped and missing income data are inadequate or waste valuable data. The next section discusses the need to develop a dependent income model, that is, a relationship between household income (the dependent variable) and a set of variables affecting household income (the independent variables), to impute a continuous income measure from grouped and missing data for use as an explanatory variable in travel demand models.

NEED FOR DISAGGREGATE INCOME MODEL FOR IMPUTING INCOME

This section discusses the need to develop a dependent income model to impute a continuous income measure. Cases in which there are no missing income data and in which there are missing income data are discussed.

No Missing Income Data

Earlier it was indicated that use of a consistent imputed estimate of income (that is, assigning to each observation falling in income category j the expected value of the income distribution bounded by the category thresholds) in a travel demand model provides consistent parameter estimates. This method assigns a single value to all income observations in a category. It does not use information on observed variables likely to affect income earnings (such as education level and number of employed adults in a household) that can help to differentiate among the incomes of different households within a particular grouped category. Developing a dependent income model (using the grouped observation on income) and combining the instrumental variable estimate of income from such a model with the information on income categories will enable construction of a consistent and efficient imputed income measure for use in travel demand models. The structure and estimation procedure for imputing income values from grouped data are discussed later in this paper.

Presence of Missing Income Data

The need to develop a dependent income model is critical when missing income data are present, since such a model is the only means of imputing an income measure for the missing data while at the same time accounting for any systematic variations in the observed characteristics (such as education level and number of employed adults) between respondent and nonrespondent households. The model should also account for systematic variations in unobserved characteristics between respondent and nonrespondent households. A consistent and efficient imputed estimate of income for use in travel demand models can be obtained from grouped and missing income data by combining the instrumental variable estimate of income from the model with information on whether a household responded to the income question or not and the income category in which a household's income falls (if the household responded). The structure and estimation procedure for imputing income values from grouped and missing income data are discussed later in this paper.

The discussion above emphasizes the need to develop a dependent income model to impute a continuous income estimate from grouped or grouped and missing income data for use as an exogenous variable in travel demand models. The remainder of this paper presents the econometric framework for imputing income through the development of a dependent income model and presents empirical results of the dependent income model and associated imputed estimates by using a Dutch data set.

ESTIMATION METHODOLOGY

The methodology used to develop a dependent income model and to impute a continuous income value from grouped and missing data in two stages is discussed in this section. In the first stage it is assumed that there are no missing income values. The methodology is then extended to accommodate missing income values in the second stage. The program routines for all estimations in this paper were written and coded by using the GAUSS matrix programming language.

No Missing Income Data

Assume that the actual but unobserved logarithm of household income, I_i^*, is a function of a vector X_i of exogenous variables as follows:

$$I_i^* = \gamma' X_i + \epsilon_i \quad (7)$$

where

γ = vector of parameters to be estimated,
X_i = vector of explanatory variables, and
ϵ_i = a random disturbance term assumed to be homoscedastic, independent, and normally distributed with mean of zero and a variance of σ^2 (a logarithm form is adopted for the dependent income variable because as indicated earlier a log-normal form has been found to be theoretically and empirically appropriate for the income distribution).

The observed data on income indicate that they fall into a prespecific interval. The relationship between the grouped observed income data I_i and the continuous unobserved (log) income value I_i^* is written as follows:

$$I_i = j \quad \text{if } a_{j-1} < I_i^* \leq a_j, \quad j = 1, \ldots J, \; i = 1, \ldots N \tag{8}$$

where the a_j's represent known threshold values (which represent the logarithm of the actual income threshold bounds) for each income category j. Representing the cumulative standard normal by Φ, the probability that household income falls in category j may be written from Equations 7 and 8 as

$$\text{Prob}(I_i = j) = \Phi\left(\frac{a_j - \gamma' X_i}{\sigma}\right) - \Phi\left(\frac{a_{j-1} - \gamma' X_i}{\sigma}\right) \tag{9}$$

Defining a set of dummy variables

$$M_{ij} = \begin{cases} 1 \text{ if } I_i^* \text{ falls in the } j\text{th category} \\ 0 \text{ otherwise,} \end{cases} \quad (i = 1, 2, \ldots N, j = 1, 2, \ldots J) \tag{10}$$

the likelihood function for estimation of the parameters γ and σ is

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[\Phi\left(\frac{a_j - \gamma' X_i}{\sigma}\right) - \Phi\left(\frac{a_{j-1} - \gamma' X_i}{\sigma}\right)\right]^{M_{ij}} \tag{11}$$

Initial parameter values for the maximum likelihood search are obtained by assigning to each income observation its conditional expectation on the basis of the marginal distribution of I^* and regressing these conditional expections on the vector of exogenous variables. The reader will note that the likelihood function of Equation 11 differs from that of the standard ordered probit model. In particular σ is unidentifiable and the threshold values (the a_j's) are unknown parameters to be estimated in the ordered probit model. In contrast in the current model the threshold values are known, and (as a consequence) σ is identifiable.

Defining the standard normal density function by $\phi(.)$, an imputed value for household (log) income may be computed for all the observations from the estimates of γ and σ obtained from maximizing the likelihood function in Equation 11. The imputed value for an income observation in category j may be computed by using the properties of doubly truncated univariate normal distributions (*12*) as follows:

$$\hat{I}_i^* | (X_i, I_i = j) = \hat{\gamma}' X_i + \hat{\sigma} \frac{\phi\left(\frac{a_{j-1,i} - \hat{\gamma}' X_i}{\hat{\sigma}}\right) - \phi\left(\frac{a_{j,i} - \hat{\gamma}' X_i}{\hat{\sigma}}\right)}{\Phi\left(\frac{a_{j,i} - \hat{\gamma}' X_i}{\hat{\sigma}}\right) - \Phi\left(\frac{a_{j-1,i} - \hat{\gamma}' X_i}{\hat{\sigma}}\right)} \tag{12}$$

These imputed values represent unbiased and consistent measures of (log) income and can be used as an explanatory variable in travel demand models (the imputed values are also guaranteed to fall within the lower and upper boundaries of the observed income categories). If an alternative function of income (other than the log function), $g(I_i^*)$, appears as the explanatory variable in the travel demand model, an imputed value may be computed as:

$$\hat{g}(I_i^*) = g(\hat{I}_i^*) \tag{13}$$

This imputed value of the function of (log) income is not unbiased, since in general the expected value of a continuous function of a variable is not equal to the function of the expected value of the variable. However it is consistent by Slutsky's theorem and thus will enable consistent estimation of travel demand models.

Presence of Missing Income Data

If missing income values are present in the data (as is almost always the case), one of two approaches may be used to construct a continuous value for all observations: (a) the naive approach or (b) the sample selection approach.

Naive Approach

The naive approach employs the method described above to estimate γ and σ by using only the observed (and grouped) income values. A continuous (log) income value is then imputed by using Equation 12 for observed income values and using $\hat{I}_i^* = \hat{\gamma}' X_i$ for missing income values. The naive approach accounts for systematic differences in the observed characteristics (represented by the X vector in Equation 7) that affect income between households that provide income and those that do not. However it fails to accommodate for systematic differences in the unobserved characteristics that affect income between respondent and nonrespondent households; that is, it ignores any "self-selection" in the choice of households to report income. Specifically unobserved factors that affect household income may also influence the decision of individuals (or households) to report income. For example it seems at least possible that households with above-average incomes, other things being equal, will be more reluctant than other households to provide information on income [Lillard et al. (*11*) indicate that this is so in their study of the 1980 Census Population Survey]. Because of this potential sample selection [see Mannering and Hensher (*13*) for a detailed review of sample selection-related issues], the naive approach will not, in general, provide consistent (continuous) estimates of income for observed or missing income data [the method proposed by Stern (*14*) for imputing income from grouped and missing income data falls under the naive approach]. To obtain consistent estimates the decision to report income should be considered endogenous, as discussed in the next section.

Sample Selection Approach

The sample selection approach uses two equations, one for income reporting and the other for household income, and accounts for the correlation in error terms between the two equations. Thus it accommodates systematic differences in unobserved characteristics between respondent and nonrespondent households. The model system is as follows:

$$r_i^* = \gamma_r' X_{ri} + \epsilon_{ri}, \; r_i = 1 \text{ if } r_i^* > 0 \text{ and } r_i = 0 \text{ if } r_i^* \leq 0 \tag{14}$$

$$\left.\begin{aligned} I_i^* &= \gamma_I' X_{Ii} + \epsilon_{Ii} \\ I_i &= j, \text{ if } a_{j-1} < I_i^* \leq a_j \end{aligned}\right\} \text{observed only if } r_i^* > 0 \qquad (15)$$

where

r_i = observed binary variable indicating whether or not income is reported ($r_i = 1$ if income is reported and $r_i = 0$ otherwise),
r_i^* = underlying continuous variable related to the observed binary variable r_i as shown above,
X_{ri} and X_{Ii} = vectors of exogenous variables,
γ_r and γ_I = vectors of parameters to be estimated, and
ϵ_{ri} and ϵ_{Ii} = normal random error terms assumed to be independent and identically distributed across observations with a mean of zero and variance of one and σ_I^2, respectively.

The error terms are assumed to follow a bivariate normal distribution (the author is not aware of any earlier application of sample selection in econometric literature in which the variable subjected to sample selection is observed only in grouped form).

The probability that income is observed and falls in income category j from the model system of Equations 14 and 15 is:

$$\text{Prob}(r_i = 1, I_i = j) = \Phi_2\left(\frac{a_j - \gamma_I' X_{Ii}}{\sigma_I}, \gamma_r' X_{ri}, -\rho\right) - \Phi_2\left(\frac{a_{j-1} - \gamma_I' X_{Ii}}{\sigma_I}, \gamma_r' X_{ri}, -\rho\right) \qquad (16)$$

where ρ is the correlation between the error terms ϵ_{ri} and ϵ_{Ii}, and Φ_2 is the cumulative standard bivariate normal function.

Defining a set of dummy variables M_{ij} as in Equation 10 for the observed income observations, the appropriate maximum likelihood function for estimation of the parameters in the model system is

$$\mathcal{L}_f = \prod_{i=1}^{N} \left[1 - \Phi(\gamma_r' X_{ri})\right]^{1-r_i} \times \left\{\prod_{j=1}^{J} \left[\Phi_2\left(\frac{a_j - \gamma_I' X_{Ii}}{\sigma_I}, \gamma_r' X_{ri}, -\rho\right) - \Phi_2\left(\frac{a_{j-1} - \gamma_I' X_{Ii}}{\sigma_I}, \gamma_r' X_{ri}, -\rho\right)\right]^{M_{ij}}\right\}^{r_i} \qquad (17)$$

Initial start values for the *ML* iterations are obtained by assigning to each reported income observation its conditional expectation on the basis of the marginal distribution of the underlying latent continuous variable I_i^*. These values are now treated as the actual continuous (log) income values, and a Heckman's two-step method (*15*) is applied for sample selection models to obtain start values for the parameters.

The continuous value of (log) income for households that reported income may be computed from the parameter estimates obtained from maximizing Equation 17. By using the properties of doubly truncated bivariate normal distributions (*16*) and defining the following quantities,

$$m = \frac{a_j - \hat{\gamma}_I' X_{Ii}}{\hat{\sigma}_I}$$

$$k = \frac{a_{j-1} - \hat{\gamma}_I' X_{Ii}}{\hat{\sigma}_I}$$

$$g = \frac{\hat{\gamma}_r' X_{ri} + k\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}$$

$$h = \frac{\hat{\gamma}_r' X_{ri} + m\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}$$

$$r = \frac{k + \hat{\gamma}_r' X_{ri}\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}$$

$$s = \frac{m + \hat{\gamma}_r' X_{ri}\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}$$

one can write

$$\hat{I}_i^* | (X_{ri}, X_{Ii}, r_i = 1, I_i = j) = \hat{\gamma}_I' X_{Ii} + \hat{\sigma}_I \frac{\phi(k)\Phi(g) - \phi(m)\Phi(h) + \hat{\rho}\phi(-\hat{\gamma}_r' X_{ri})[\Phi(s) - \Phi(r)]}{\Phi_2(\hat{\gamma}_r' X_{ri}, m, -\hat{\rho}) - \Phi_2(\hat{\gamma}_r' X_{ri}, k, -\hat{\rho})} \qquad (18)$$

The above expression collapses to Equation 12 if the correlation between the error terms in the reporting equation and the income equation is zero.

The continuous value of (log) income for households that did not report income may be imputed as follows:

$$\hat{I}_i^* | (X_{ri}, X_{Ii}, r_i = 0) = \hat{\gamma}_I' X_{Ii} - \hat{\rho}\hat{\sigma}_I \left(\frac{\phi(\hat{\gamma}_r' X_{ri})}{1 - \Phi(\hat{\gamma}_r' X_{ri})}\right) \qquad (19)$$

EMPIRICAL RESULTS

This section discusses the data used to develop the dependent income model and to impute income from grouped and missing income observations and also presents estimation results.

Data

The data source used in the present study is from a Dutch National Mobility Survey. The survey involved weekly travel diaries and household and personal questionnaires collected during the spring of 1988 [for a detailed description of this survey see van Wissen and Meurs (*17*)]. The sample included 889 households, 55 of which have missing income data. Household income was available in three categories (for the observed income observations) in the data: (a) less than or equal to 24,000 guilders, (b) from 24,001 to 38,000 guilders, and (c) greater than 38,000 guilders.

Empirical Specification and Results

The variables considered in the income reporting equation and household income equation are listed in Table 1. They included household age and education (see definitions in Table 1), number of employed adults in the household, number of kids in the house-

TABLE 1 Exogenous Variables in Model

Variable	Definition
household age	average age of adults in the household (age of adult in single adult households)
household age > 35	(household age-35) if household age greater than 35, 0 otherwise
household age > 45	(household age-45) if household age greater than 45, 0 otherwise
household secondary education	1 if education of all adults in the household is at secondary level, 0 otherwise
household high-secondary education	1 if education of adults in household is a mixture of high and secondary, 0 otherwise
household high education	1 if education of all adults in the household is high, 0 otherwise
number of employed adults	number of employed adults in the household
number of kids	number of children less than 12 in household
returning young adult (RYA) family	1 if household has a returning young adult 0 otherwise
unemployment rate	unemployment rate in the municipality of household residence

Note: The base for the household education variables is household primary education; that is, households with one or more adults with primary education.

hold, an indicator of whether the household has a "returning" young adult, and unemployment rate in the municipality of household residence. The household age variables enable nonlinear estimation of the age effect on income reporting and income earnings. The education variables indicate the effect of different levels of education of the adults in the household relative to that for households with one or more adults with primary education.

The naive method and the sample selection method were used to estimate the parameters in the household income equation. The naive method estimates parameters from observed income observations by using Equation 11, whereas the sample selection method estimates parameters from all observations by using Equation 17. The results are shown in Table 2. The naive method estimates only the income equation, whereas the sample selection method estimates both the reporting equation and the income equation and accounts for the correlation in unobserved factors that affects these equations simultaneously. In both models the level of household education and the number of employed adults have a positive effect on income. The magnitudes of the parameters on household education are consistent with the expectation that higher levels of education have a greater effect on income. The unemployment rate in the municipality of the household residence has a significant negative effect. The reporting equation estimation results in the sample selection model indicate that households with older adults, households whose individuals have a high level of education, and households with a returning young adult have a significant negative effect on reporting. Thus there are systematic differences in the observed characteristics between households that report income and those that do not.

The magnitude and significance of the correlation term ρ in the sample selection model indicate that there is a significant (and rather high) negative correlation in the unobserved factors that affect the reporting equation and the household income equation; that is, households that did not report their incomes were, all observed characteristics being equal, likely to have higher incomes than households that reported their incomes. This indicates that the naive method provides biased and inconsistent estimation results. In particular the naive method tends to underestimate the magnitudes of parameters on the exogenous variables that have a positive effect on income and tends to overestimate the magnitudes of parameters on the exogenous variables that have a negative effect on income in the income equation because of the negative correlation between the error terms in the reporting and the income equations (although the difference in coefficient estimates between the naive and the sample selection approaches appears to be small, the reader should note that the dependent variable is the logarithm of income, and thus even small coefficient differences could translate into moderate differences with respect to income earnings; the small coefficient differences may also be attributable to the small number of missing income observations in the current data set).

The mean values of imputed (log) income for households that reported income and those that did not report income obtained by using the midpoint method, the naive method, and the sample selection method are shown in Table 3. The mean values for the midpoint method depend on the representative value used for the lowest and the highest income categories. In the computations shown in Table 3 a value of log (15,000) was assigned for the

TABLE 2 Estimation Results

Equation	Variable	The naive approach		The sample selection approach	
		Coefficient	t stat.	Coefficient	t stat.
Reporting equation	constant	-	-	2.218	1.77
	household age				
	entire range	-	-	0.002	0.04
	> 35 years	-	-	0.065	0.90
	> 45 years	-	-	-0.141	-2.31
	household education				
	secondary/high	-	-	-0.762	-3.50
	high	-	-	-1.114	-4.54
	number of kids	-	-	-0.150	-1.27
	RYA family	-	-	-0.759	-2.08
Income equation	constant	10.053	57.35	10.051	56.75
	household age				
	entire range($x10^{-1}$)	-0.006	0.12	-0.002	0.05
	> 35 years	0.012	1.54	0.011	1.42
	> 45 years	-0.006	-0.94	-0.004	-0.62
	household education				
	secondary	0.188	6.63	0.188	6.60
	secondary/high	0.364	10.67	0.382	11.05
	high	0.414	10.12	0.446	10.59
	number of employed adults	0.258	10.83	0.258	10.73
	unemployment rate	-1.125	-3.20	-1.117	-3.19
	σ	0.267	19.36	0.273	18.15
Correlation term	ρ	-	-	-0.694	-2.33
# of observations		834		889	
Log Likelihood (slopes=0, ρ=0)		-800		-1006	
Log Likelihood (convergence)		-621		-797	

TABLE 3 Mean Values of Imputed Income

Category	Mean imputed (log) income		
	Midpoint Approach	Naive Approach	Sample Selection Approach
Households which reported income (respondent households)	10.419 (0.319)	10.550 (0.245)	10.528 (0.317)
Households which did not report income (non-respondent households)	10.419 (0.000)	10.713 (0.280)	11.036 (0.240)

Note: Numbers in parentheses are centered standard deviations.

"less than or equal to 24,000 guilders" category and a value of log (43,000) was assigned for the "greater than 38,000 guilders" category. The inconsistency and the ad hoc nature of the midpoint method of imputing income were discussed above. Furthermore the mean value of imputed (log) income was identical for both respondent and nonrespondent households by the midpoint method because the midpoint method does not account for systematic variations in the observed and unobserved characteristics that affect income between respondent and nonrespondent households.

The naive method accounts for systematic variations in observed characteristics between respondent and nonrespondent households. The higher mean estimate for nonrespondent households compared with that for respondent households indicates that nonrespondent households have higher values than respondent households for the observed characteristics that increase income. This is readily observed in the reporting equation estimates of the sample selection model in Table 2, which indicate that nonrespondent households are characterized by adult members with a higher education level than those of adult members in respondent households.

The sample selection method accounts for systematic variations in the observed and unobserved characteristics that affect income between respondent and nonrespondent households. The difference in the mean value of imputed (log) income for respondent and nonrespondent households between the sample selection and naive approaches comprises two components. The first component is an underestimation of income by the naive method on the basis of the observed characteristics that affect income because of the biases in parameter estimates of the naive approach in Table 2. This first component leads to an increase in imputed (log) incomes for both respondent and nonrespondent households in the sample selection method compared with those in the naive method. The second component is the effect of the unobserved characteristics that affect reporting status and income. It leads to a decrease in imputed (log) income for respondent households and an increase for nonrespondent households. The naive method does not consider this second component; only the sample selection model does. The difference in the mean value of imputed (log) income between the sample selection and naive approaches is small for respondent households because the two components mentioned above act in opposite directions and tend to offset each other. On the other hand the mean value of imputed (log) income from the sample selection approach is substantially larger than that from the naive approach for nonrespondent households because the two components mentioned above reinforce each other. Aside from the magnitude of the difference between the estimates of the sample selection and the naive approaches, however, the naive approach provides inconsistent imputed estimates both for respondent and for nonrespondent households because the correlation in the unobserved factors that affect reporting status and income earnings is significantly different from zero in Table 2. In general the sample selection method is the only approach that provides consistent imputed income estimates from grouped and missing income data.

CONCLUSION

This paper developed a methodology for imputing a continuous value of income from grouped and missing income data for use as an explanatory variable in travel demand models. The method was applied to data from the Dutch National Mobility Survey. In addition to indicating the applicability of the procedure developed in the paper to accommodating grouped and missing data, the results show that there are systematic differences in observed and unobserved characteristics between households that report income and households that do not. Failure to accommodate for this sample selection results in biased and inconsistent imputations. Use of such inconsistent imputed income values as an explanatory variable will result in unreliable travel demand models.

The methodology developed in this paper is particularly relevant because almost all transportation-related data bases record income in grouped form and because there is a trend for an increasing percentage of respondents to refuse to provide income information in travel and travel-related surveys (*11*). The methodology developed in the paper is easy to apply and has been coded for use with the GAUSS programming language.

ACKNOWLEDGMENTS

The author would like to thank Frank Koppelman and three anonymous referees for useful suggestions on previous versions of this paper.

REFERENCES

1. Golob, T. F. The Dynamics of Household Travel Time Expenditures and Car Ownership Decisions. Presented at the International Conference on Dynamic Travel Behavior Analysis, Kyoto, Japan, July 1989.
2. Meurs, H. Dynamic Analysis of Trip Generation. Presented at the International Conference on Dynamic Travel Behavior Analysis, Kyoto, Japan, July 1989.
3. Beggan, J. G. *The Relationship Between Travel/Activity Behavior and Mode Choice for the Work Trip.* M.S. thesis. Transportation Center, Northwestern University, Evanston, Ill., 1988.
4. Stewart, M. B. On Least Squares Estimation When the Dependent Variable Is Grouped. *Review of Economic Studies*, 1983, pp. 737–753.
5. Churchill, G. A., Jr., *Marketing Research: Methodological Foundations.* The Dryden Press, Chicago, 1983.
6. Hamburg, J. R., E. J. Kaiser, and G. T. Lathrop. *NCHRP Report 266: Forecasting Inputs to Transportation Planning.* TRB, National Research Council, Washington, D.C., 1983.
7. Hsiao, C. Regression Analysis with a Categorized Explanatory Variable. In *Studies in Econometrics, Time Series, and Multivariate Statistics.* Academic Press, Incorporated, New York, 1983.
8. Aitchison, J., and J. A. C. Brown. *The Lognormal Distribution with Special Reference to Its Uses in Economics.* Cambridge University Press, Cambridge, 1976.
9. Mincer, J. *Schooling, Experience and Earnings.* National Bureau of Economic Research, New York, 1974.
10. Haitovsky, Y. *Regression Estimation from Grouped Observations.* Hafner Press, New York, 1973.
11. Lillard, L., J. P. Smith, and F. Welch. What Do We Really Know About Wages? Importance of Nonreporting and Census Information. *Journal of Political Economy*, Vol. 94, No. 31, 1986, pp. 489–506.
12. Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distribution.* John Wiley & Sons, Incorporated, New York, 1972.
13. Mannering, F., and D. A. Hensher. Discrete/Continuous Econometric Models and their Applications to Transport Analysis. *Transport Reviews*, Vol. 7, No. 3, 1987, pp. 227–244.
14. Stern, S. Imputing a Continuous Income Variable from a Bracketed Income Variable with Special Attention to Missing Observations. *Economic Letters*, Vol. 37, 1991, pp. 287–291.

15. Heckman, J. J. Sample Selection Bias as a Specification Error. *Econometrica*, Vol. 47, 1979, pp. 153–161.
16. Shah, S. M., and N. T. Parikh. Moments of Singly and Doubly Truncated Standard Bivariate Normal Distribution, *Vidya*, Vol. 7, 1964, pp. 82–91.
17. van Wissen, L., and H. J. Meurs. The Dutch National Mobility Panel: Experiences and Evaluation. *Transportation*, Vol. 16, No. 2, 1989.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.