ELSEVIER

Economics Letters 46 (1994) 311–319

economics letters

Imputing a continuous income variable from grouped and missing income observations

Chandra R. Bhat

235 Marston Hall, Department of Civil Engineering, University of Massachusetts, Amherst, MA 01003, USA

Received 13 October 1993; final revision received 29 December 1993; accepted 20 April 1994

Abstract

Most cross-sectional data sets collect income in a discrete number of categories (that is, 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. This paper develops a method to impute a continuous and reliable value for income from grouped and missing income data.

JEL classification: C34

1. Introduction

In many cross-sectional data sets, 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 response to the income question is also frequently associated with high non-response rates, leading to missing income observations.

While income is measured in grouped form, it is the continuous measure of income that frequently appears as an explanatory variable in labor supply models, market research models and travel demand models (Killingsworth, 1983; Koppelman et al., 1993; Golob, 1989). Alternatively, a researcher may want to use a continuous measure to conserve on degrees of freedom. A common procedure to handle grouped and missing income data is to assign the midpoint of the known income threshold bounds determining each category to observations in that category (an arbitrary truncation point is used as the representative value for the two categories at either end of the income spectrum) for observed income observations and to drop all the missing income observations or assign the average value of the midpoint estimates of the observed income observations to the missing income observations (we will refer to this procedure as the midpoint approach). Hsiao (1983) indicates that assigning midpoints of categories to observations in that category and using the resulting continuous income variable as an explanatory variable in a model results in inconsistent model relationships. Dropping all

0165-1765/94/$07.00
SSDI 0165-1765(94)00501-R

missing income observations also has its problems. If systematic variations in income level are present between respondent and non-respondent households (or individuals), then the model relationship for non-respondents may be different from that of respondents. Thus, a model relationship obtained by dropping non-respondents will not be a representative relationship for the entire population. Also, dropping missing income observations results in loss of observations. Finally, the alternative of assigning the average value of observed income observations to missing income observations assumes that the average income of respondents is identical to that of non-respondents. This is not likely to be the case because of systematic variations in observed and unobserved characteristics (e.g. education, sensitivity to privacy, etc.) affecting income earnings between respondents and non-respondents.

This paper proposes a method to construct a continuous measure of income for all observations in a data set with grouped and missing income data. Section 2 discusses the work of Stewart (1983) and Stern (1991), which motivates the procedure developed in the remainder of the section. Section 3 presents empirical results. Section 4 provides a summary of the research and highlights important findings.

2. Previous imputation methods and proposed methodology

Stewart (1983) developed a model with income as the dependent variable considering that grouped income was available for all observations. His model system is as follows:

$$I_i^* = \gamma_I' X_{Ii} + \epsilon_{Ii} ,$$
$$I_i = j , \quad \text{if } a_{j-1} < I_i^* \leq a_j , \tag{1}$$

where I_i^* is the true (but unobserved) logarithm of income, X_{Ii} is a vector of explanatory variables, ϵ_{Ii} is a random disturbance term assumed to be homoscedastic, independent, and normally distributed with mean 0 and variance σ_I^2, γ_I is a vector of parameters to be estimated, I_i is grouped observed income, and the a_j's represent known threshold values for each income category j. Representing the cumulative standard normal by Φ and defining a set of dummy variables

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

the likelihood function for estimation of the parameters γ_1 and σ_I is

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[\Phi\left(\frac{a_j - \gamma_I' X_{Ii}}{\sigma_I}\right) - \Phi\left(\frac{a_{j-1} - \gamma_I' X_{Ii}}{\sigma_I}\right) \right]^{M_{ij}} . \tag{3}$$

An unbiased and consistent measure of (log) income may then be imputed for an income observation in category j as follows:

$$\hat{I}_i^* | (X_{Ii}, I_i = j) = \hat{\gamma}_I' X_{Ii} + \hat{\sigma}_I \frac{\phi\left(\frac{a_{j-1,i} - \hat{\gamma}_I' X_{Ii}}{\hat{\sigma}_I}\right) - \phi\left(\frac{a_{j,i} - \gamma_{\hat{I}}' X_{Ii}}{\hat{\sigma}_I}\right)}{\Phi\left(\frac{a_{j,i} - \hat{\gamma}_I' X_{Ii}}{\hat{\sigma}_I}\right) - \Phi\left(\frac{a_{j-1,i} - \hat{\gamma}_I' X_{Ii}}{\hat{\sigma}_I}\right)}. \tag{4}$$

A procedure to handle missing income observations within Stewart's framework is to maximize (3) using observed income observations. Eq. (4) can then be used to impute a continuous measure for observed income observations, while we can use the expression $\hat{I}_i^* = \hat{\gamma}_I' X_{Ii}$ to impute continuous values for missing income observations.

Stern (1991) adopts a procedure very similar to the one discussed above. He uses observations for which grouped income is observed to develop a relationship between a continuous transformed income variable and explanatory variables, employing a standard ordinal probit method. This method involves the estimation of the a_j's in Eq. (3) with $\sigma_I = 1$. The a_j's are unknown thresholds on the transformed income scale ($a_i = -\infty$, $a_2 = 0$, and $a_J = +\infty$). Continuous values of income on this transformed scale are subsequently computed from Eq. (4) (with $\sigma_I = 1$) for observed income observations. For missing income observations, the continuous value is computed as $\hat{\gamma}_I' X_{Ii}$. These continuous values on the transformed scale are transposed into a continuous value of (log) income by assuming a linear spline correspondence between the known category thresholds on the (log) income scale and the estimated a_j's.

The procedures discussed above to account for grouped and missing income data (which we shall refer to as the naive approach) fail to accommodate for systematic differences in unobserved characteristics affecting income between respondent and non-respondent households (or individuals); that is, they ignore any 'self-selection' in the choice of households to report income. Specifically, unobserved factors that affect household income may also influence the decision of households to report income. For example, it seems at least possible that households with above-average income, other things being equal, will be more reluctant than other households to provide information on income (Lillard et al., 1986, indicate that this is so in their study of the 1980 Census Population Survey). Due to this potential sample selection, the naive approach will not, in general, provide unbiased and consistent estimates of income both for observed and missing income observations. The decision to report income should be considered endogenous to obtain consistent estimates, as I discuss next.

The model system I propose (which we shall refer to as the sample selection approach) comprises two equations, one for reporting (whether income is reported or not) and the other for income earnings, and accounts for the correlation in error terms between the two equations. The model system is as follows:

$$\left.\begin{aligned} & r_i^* = \gamma_r' X_{ri} + \epsilon_{ri}\,, \quad r_i = 1 \text{ if } r_i^* > 0 \text{ and } r_i = 0 \text{ if } r_i^* \leq 0\,, \\ & I_i^* = \gamma_I' X_{Ii} + \epsilon_{Ii} \\ & I_i = j\,, \quad \text{if } a_{j-1} < I_i^* \leq a_j \end{aligned}\right\} \text{observed only if } r_i^* > 0\,, \tag{5}$$

where r_i is the observed binary variable indicating whether or not income is reported ($r_i = 1$ if income is reported and $r_i = 0$ otherwise), r_i^* is an underlying continuous variable related to the observed binary variable r_i as shown above, X_{ri} is a vector of exogenous variables influencing

the reporting decision, ϵ_{ri} and ϵ_{Ii} are 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. All other notation is as defined earlier. The probability that income is observed and falls in income category j is

$$\text{Prob}(r_i = 1, I_i = j) = \Phi_2\left(\frac{a_j - \gamma_I' X_{Ii}}{\sigma_I}, \gamma_r' X_{ri}, -\rho\right) - \Phi_2\left(\frac{a_{j-1} - \gamma_I' X_{Ii}}{\sigma_I}, \gamma_r' X_{ri}, -\rho\right), \tag{6}$$

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 Eq. (2) for the observed income observations, the appropriate maximum likelihood function for estimation of the parameters in the model system is[1]

$$\mathcal{L}_f = \prod_{i=1}^{N} [1 - \Phi(\gamma_r' X_{ri})]^{1-r_i} \times \left[\prod_{j=1}^{J} \left\{\Phi_2\left(\frac{a_j - \gamma_I' X_{Ii}}{\sigma_I}, \gamma_r' X_{ri}, -\rho\right) - \Phi_2\left(\frac{a_{j-1} - \gamma_I' X_{Ii}}{\sigma_I}, \gamma_r' X_{ri}, -\rho\right)\right\}^{M_{ij}}\right]^{r_i}. \tag{7}$$

The program routine for maximization of the above function was written and coded using the GAUSS matrix programming language. The continuous value of (log) income for households which reported income may be computed from the parameter estimates obtained from maximizing Eq. (7). Using the properties of doubly truncated bivariate normal distributions (Shah and Parikh, 1964) and defining the following quantities:

$$m = \frac{a_j - \hat{\gamma}_I' X_{Ii}}{\hat{\sigma}_I}, \quad k = \frac{a_{j-1} - \hat{\gamma}_I' X_{Ii}}{\hat{\sigma}_I}, \quad g = \frac{\hat{\gamma}_r' X_{ri} + k\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}, \quad h = \frac{\hat{\gamma}_r' X_{ri} + m\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}},$$

$$r = \frac{k + \hat{\gamma}_r' X_{ri}\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}} \quad \text{and} \quad s = \frac{m + \hat{\gamma}_r' X_{ri}\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}},$$

we can write

$$\hat{I}_i^* \mid (X_{ri}, X_{Ii}, r_i = 1, I_i = j) = \hat{\gamma}_I' X_{Ii} + \hat{\sigma}_I \frac{\phi(k)\Phi(g) - \phi(m)\Phi(h) + \hat{\rho}\phi(-\hat{\gamma}_r' X_{ri})[\Phi(s) - \Phi(r)]}{\Phi_2(\hat{\gamma}_r' X_{ri}, m, -\hat{\rho}) - \Phi_2(\hat{\gamma}_r' X_{ri}, k, -\hat{\rho})}. \tag{8}$$

The expression above guarantees that the predicted value of income for an observation in category j is within the threshold bounds a_{j-1} and a_j. In the special case that the correlation in error terms between the reporting equation and the income equation, ρ, is zero, the above expression collapses to Eq. (4).

[1] I am not aware of any application of this variant of sample selection in the econometric literature.

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

$$\hat{I}_i^* \mid (X_{ri}, X_{Ii}, r_i = 0) = \hat{\gamma}_I' X_{Ii} - \hat{\rho}\hat{\sigma}_I \left(\frac{\phi(\hat{\gamma}_r' X_{ri})}{1 - \Phi(\hat{\gamma}_r' X_{ri})} \right). \tag{9}$$

3. Empirical results

The sample selection method discussed in Section 2 is applied to data from the 1990 U.S. Nationwide Personal Transportation Survey. This survey (the reader is referred to documentation by Research Triangle Institute, 1991, for additional details about the survey) involved weekly travel diaries and household and personal questionaires, including information on annual income. Annual income from the survey[2] was artificially grouped into three categories: (a) less than \$15,000, (b) \$15,000–29,999, and (c) greater than or equal to \$30,000; for the present empirical study. The sample used includes 2136 single-individual households, 497 of whom (selected from the high-income ranges to create a sample selection bias) were assumed to have missing grouped income information.

The variables considered in the income reporting equation and the earnings equation are listed in Table 1. The age variables enable non-linear estimation of the age effect. The education variables indicate the effect of different levels of education relative to that of primary education (one or more years of pre-college schooling).

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 using Eq. (3), while the sample selection method estimates parameters from all observations using Eq. (7). In addition, we also estimated the parameters from a linear regression using the actual continuous (log) income data and from Eq. (3) using the grouped income categories for all observations *before* artificially partitioning the data into available and missing income observations (this corresponds to the case of grouped income data, but no missing income observations; we will refer to this estimation as the unpartitioned grouped income estimation). The results are shown in Table 2. All models indicate a non-linear age effect on income earnings; age has a positive effect till age 35, but has a net negative effect (computed as the sum of the coefficients on *AGE* and *AGE35*) beyond age 35. As expected, employment status, male gender, the indicator for urban area status, and education have positive effects on income. Non-caucasians have a lower income than caucasians. Finally, the census region dummy variables indicate a lower income in the north central/west and south parts of the country relative to the north eastern region.

Comparing the estimates from the different approaches, we observe that the parameters in the linear regression and unpartitioned grouped estimations are closed to one another. Between the naive and sample selection approaches, the sample selection estimates are closer to the unpartitioned grouped income and linear regression estimates. The reporting equation estimates in the sample selection estimation were as follows:

[2] Data on income is collected in 17 finely grouped categories in the survey. We assume that the midpoint of each income category represents continuous income for observations in that category. This is defendable because of the very fine categorization of income.

Table 1
List of exogenous variables in model

Variable	Definition
AGE	Age of individual
AGE35	(Age–35) if age greater than 35, 0 otherwise
EMPL	1 if individual is employed, 0 otherwise
MALE	1 if individual is male, 0 otherwise
URBAN	1 if individual resides in urban area, 0 otherwise
SECEDUC	1 if individual has had undergraduate education but no graduate education, 0 otherwise
HIEDUC	1 if individual has had graduate education, 0 otherwise
NONCAUCS	1 if individual is not a Caucasian 0 otherwise
AFRICAN	1 if individual is an African American 0 otherwise
SOUTH	1 if individual residence is in South Census region, 0 otherwise
NORCENWEST	1 if individual residence is in North Central or West Census regions, 0 otherwise
NORCENSOUTH	1 if individual residence is in North Central or South Census regions, 0 otherwise

Note: The base for the education variables is primary education; that is, one or more years of pre-college schooling.

$$\hat{r}_i^* = \underset{(8.06)}{3.820} \underset{(-5.30)}{- 0.073\, AGE_i} + \underset{(5.07)}{0.083\, AGE35_i} \underset{(-3.84)}{- 0.687\, EMP_i} \underset{(-4.49)}{- 0.518\, SECEDUC_i}$$
$$\underset{(-5.68)}{- 0.941\, HIEDUC_i} + \underset{(2.10)}{0.351\, AFRICAN_i} + \underset{(2.59)}{0.284\, NORCENSOUTH_i}\,.$$

Numbers in parenthesis below coefficient estimates are t-statistics. The reason for the better performance of the sample selection approach is that it considers income reporting to be

Table 2
Income equation estimation results

Variables	Linear regression		Unpartitioned grouped estimation		The naive approach		The sample selection approach	
	Coefficient	*t* stat.	Coefficient	*t* stat.	Coefficient	*t* stat.	Coefficient	*t* stat.
Constant	7.692	62.66	8.090	61.80	8.700	85.04	8.470	53.72
AGE	0.044	11.81	0.042	10.62	0.024	7.97	0.033	6.60
AGE35	−0.050	−10.85	−0.049	−10.29	−0.031	−8.31	−0.041	−6.98
EMPL	0.684	17.66	0.518	11.74	0.368	10.93	0.440	8.83
MALE	0.177	5.85	0.171	5.73	0.074	2.96	0.076	2.96
URBAN	0.164	5.07	0.110	3.26	0.054	1.96	0.060	2.01
SECEDUC	0.323	9.73	0.333	10.03	0.206	7.39	0.272	6.47
HIEDUC	0.540	11.89	0.511	10.63	0.233	5.65	0.364	5.04
NONCAUCS	−0.192	−4.82	−0.171	−4.50	−0.083	−2.70	−0.109	−3.14
NORCENWEST	−0.077	−2.24	−0.089	−2.58	−0.075	−2.55	−0.097	−3.03
SOUTH	−0.192	−4.85	−0.210	−5.54	−0.148	−4.66	−0.184	−4.91
Standard error	0.658	–	0.564	31.90	0.405	32.31	0.450	12.21
Correlation	Not applicable		Not applicable		Not considered		−0.622	−3.24

endogenous and accounts for the correlation between unobserved factors affecting reporting status and income earnings. This correlation in unobserved factors is negative, high in magnitude, and significant, as shown in Table 2 in the final row of the sample selection column. This indicates that individuals who withheld reporting their income were, all observed characteristics being equal, likely to have higher incomes than households that reported their incomes. Thus the sample selection method correctly identifies and accommodates the sample selection bias in the partitioned data.

Table 3 indicates the mean square error (MSE) of imputed income values (relative to the

Table 3
Goodness of fit of income imputations[a]

Sample	Unpartitioned grouped estm.[b]	Midpoint approach[c]	Naive approach	Sample selection approach
Individuals assumed to have reported income (observed income sample)	0.176	0.204	0.195	0.187
Individuals assumed to have witheld income information (missing income sample)	0.096	1.220	0.844	0.264
Overall Sample	0.157	0.440	0.346	0.204

[a] Goodness of fit is measured as the mean squared error relative to the actual continuous income values.
[b] Refers to income imputations obtained from grouped data when no missing income observations are assumed.
[c] A value of log(10,000) is assigned to all observations in the 'less than $15,000' category and a value of log(45,000) is assigned to all observations in the 'greater than or equal to $30,000' category. All missing income observations are assigned a value equal to the average of the imputed income values for the observed income observations.

actual continuous income values) for the midpoint, naive and sample selection approaches.[3] We have also computed the mean square error for the imputed values resulting from the unpartitioned grouped estimation, which represents the minimum achievable error in the presence of grouped and missing income observations. Thus it serves as a yardstick to evaluate the performance of the midpoint, naive, and sample selection approaches. As observed from Table 3, the MSE from the three approaches are close to the MSE for the unpartitioned estimation for individuals who were assumed to have reported their income (observed income sample), with the MSE for the sample selection method being closest and the MSE for the midpoint method being farthest. However, the MSE for the midpoint and naive methods are high for individuals who were assumed not to report their income (missing income sample), while the MSE for the sample selection method is much more reasonable. The MSEs for the overall sample are provided in the final row. The MSE for the sample selection method is only 30% higher than that for the unpartitioned case compared with 180% higher for the midpoint method and 120% higher for the naive method. This is a clear indication that the sample selection method developed in this paper is the preferred approach when imputing continuous income values from grouped and missing income observations.

4. Conclusion

This paper has developed a methodology to impute a continuous value of income from grouped and missing income data, accounting for sample selection in income based on the decision to report income. The method is easy to apply and has been coded for use with the GAUSS programming language. The method was applied to data from the 1990 Nationwide Personal Transportation Survey. The results, in addition to indicating the applicability of the procedure developed in the paper to accommodate grouped and missing data, show that the procedure provides much better income imputations compared with the midpoint or the naive approaches. However, it should be emphasized that this conclusion is specific to the situation where (a) the income intervals used in data collection are broad and (b) there is a sizeable number of missing income observations. If there are relatively few missing income observations, the naive method is likely to provide reasonably accurate income imputations. If, in addition, the income intervals used in data collection are very fine, the midpoint method may suffice to provide accurate income imputations.

References

Golob, T.F., 1989, The dynamics of household travel time expenditures and car ownership decisions, presented at the International Conference on Dynamic Travel Behavior Analysis, Kyoto, Japan, July.

[3] We assigned a value of log(10,000) for the 'less than $15,000' category and a value of log(45,000) for the 'greater than or equal to $30,000' category for the midpoint method computations. All missing income observations are assigned a value equal to the average of the imputed values for the observed income observations.

Hsiao, C., 1983, Regression analysis with a categorized explanatory variable, in: Studies in econometrics, time series, and multivariate statistics (Academic Press, New York).

Killingsworth, M.R., 1983, Labor supply (Cambridge University Press, Cambridge).

Koppelman, F.S., C.R. Bhat and J.L. Schofer, 1993, Market research evaluations of actions to reduce suburban traffic congestion: Commuter travel behavior and response to demand reduction actions, Transportation Research 27A, no. 5, 383–393.

Lillard, L., J.P. Smith and F. Welch, 1986, What do we really know about wages? The importance of nonreporting and census information, Journal of Political Economy 94, no. 31, 489–506.

Research Triangle Institute, 1991, 1990 nationwide personal transportation survey: User's guide to the public use of tapes, submitted to the Department of Transportation, Federal Highway Administration, December.

Shah, S.M. and N.T. Parikh, 1964, Moments of singly and doubly truncated standard bivariate normal distribution, Vidya 7, 82–91.

Stern, S. 1991, Imputing a continuous income variable from a bracketed income variable with special attention to missing observations, Economic Letters 37, 287–291.

Stewart, M.B., 1983, On least squares estimation when the dependent variable is grouped, Review of Economic Studies, 737–753.