IMPROVING INCOME IMPUTATION BY USING PARTIAL INCOME INFORMATION AND ECOLOGICAL VARIABLES

Michael P. Battaglia, David C. Hoaglin, David Izrael, Abt Associates Inc.; Meena Khare, National Center for Health Statistics; and Ali Mokdad, Centers for Disease Control and Prevention

Michael P. Battaglia, Abt Associates Inc., 55 Wheeler St., Cambridge, MA 02138

Key Words: Random-Digit Dialing, Income Cascading, National Immunization Survey

Introduction

Many household telephone surveys collect family income. It is sometimes used as a poststratification variable for the survey weights. It is more often used as a classification variable in analyses of survey outcome variables (e.g., health status measures). Family income, along with the number of persons in the family and the number of children in the family, determines poverty status. Family income can also be divided by the appropriate poverty threshold to form an income-to-poverty ratio in order to quantify different poverty levels (e.g., 100%, 150% and 200% of poverty).

The collection of family income in telephone surveys is not without problems. It is usually not possible to ask the respondent to provide a detailed breakdown of income by source or to consult records. Item nonresponse on family income can range from 15% to 30% or even higher. The high level of item nonresponse encountered in most surveys makes it difficult to use family income for weighting or analysis. For example, in showing estimates for a key survey variable by poverty status, it is necessary to show three categories: below poverty level, at or above poverty level, and poverty level unknown. Our research focused on imputing missing family incomes in the National Immunization Survey (NIS), a telephone survey, and using the resulting information to create imputed poverty status and income-to-poverty ratio variables. We used the wording structure of the NIS income questions to construct groups corresponding to several levels of partial information on income and also employed a series of ecological variables in imputing family income.

The National Immunization Survey

The NIS is an ongoing telephone survey of children aged 19-35 months sponsored by the Centers for Disease Control and Prevention. Each quarter a stratified sample of telephone numbers is selected in 78 geographic areas, and the released sample is called to screen for households with age-eligible children. Once an eligible household is identified, the person most knowledgeable about the child(ren)'s vaccinations is interviewed. Approximately 35,000 interviews are completed each year. The questionnaire asks about the vaccinations received by the child and then goes on to collect a series of background variables including family income in the last calendar year (e.g., income in 1999 for the 2000 NIS).

The household interview includes several questions about family income (Olson et al., 1999). These questions ask first for the amount of total family income from all sources in the past calendar year. The interviewer enters the reported income, and the Computer Assisted Telephone Interviewer (CATI) system then parses the response for the interviewer to read back to the respondent, to confirm that the family income was recorded correctly (e.g., "83 thousand dollars"). If the respondent does not know or refuses to answer this question, the interview continues with a cascading sequence of income questions that would place the family income in one of a set of intervals. The cascading entry point is at $20,000:

You may not be able to give us an exact figure for your total combined family income, but was your total family income during (LAST CALENDAR YEAR) more or less than $20,000?

MORE THAN $20,000 1
$20,000..2
LESS THAN $20,000 3
DON'T KNOW..............................6
REFUSED.....................................7

A total of 15 cascading questions attempt to place the family income into one of 15 income intervals:

1.	$0	-	$7,500
2.	$7,501	-	$10,000
3.	$10,001	-	$12,500
4.	$12,501	-	$15,000
5.	$15,001	-	$17,500
6.	$17,501	-	$20,000
7.	$20,001	-	$25,000
8.	$25,001	-	$30,000
9.	$30,001	-	$35,000
10.	$35,001	-	$40,000
11.	$40,001	-	$45,000
12.	$45,001	-	$50,000
13.	$50,001	-	$60,000
14.	$60,001	-	$75,000
15.	$75,001	or	greater

The resulting composite variable places each child into one of these 15 intervals or into one of three missing-value categories: don't know, refused, or no answer given.

Item Nonresponse

In the 2000 NIS 27.8% of respondents did not answer the question about the total family income in the past calendar year. Among those nonrespondents, 51.1% completed the income cascading questions, resulting in an item nonresponse rate of 14.2% for the family income composite variable.

Among the 14.2% with a missing value on the family income composite variable, 39.1% of respondents completed part of the income cascading questions. Although these children are missing on the family income composite variable, the partial income information does locate their income in an interval (a union of two or more adjacent intervals among the 15), and can be used to improve imputation of family income. The imputation methods studied include regression models whose sets of predictor variables include demographic and socioeconomic characteristics, and telephone-exchange-level ecological variables. As discussed below, some use was also made of hot-deck imputation.

As shown in Table 1, the response to the total family income question and the partial income cascading information lead to 16 groups of children who have a missing value on the family income composite variable. Group 1 consists of children with no partial income information (i.e., they did not fall into Groups 2-16). This group consists primarily of interview break-offs before the family income questions. For Groups 2-7 income is known to within $20,000 or is located in an open-ended interval. Group 8 children have a "DK" response to the total family income question, and no cascading information is available. Group 9 children have a "Refused" response to the total family income question, and no cascading information was obtained. Finally, for Groups 10-16 income is known to within $10,000 or within $5,000.

Among the 14.2% of children with a missing composite family income variable, 24.2% are in Groups 2 to 7. The "DK" group accounts for 30.9%, and the "Refused" group accounts for another 21.4%. The groups where the income is known to within $10,000 or within $5,000 account for 14.9%. Group 1, children with no partial income information, accounts for 8.7% of the children with a missing family income composite variable.

As mentioned earlier, about half of the respondents who gave "DK" or "Refused" on total family income did complete the cascading questions. The resulting data allowed us to compare the distribution of family income between those who responded "DK" and those who refused. We found that 57.1% of "DK" respondents who completed the cascade had family income less than or equal to $20,000, compared to 15.2% for respondents who refused. Only 18.4% of those "DK" respondents had family income greater than $50,000, compared to 57.2% for those who refused. Also, the children with a "Refused" response tended to have demographic and socioeconomic characteristics associated with higher-income families, compared to the children with a "DK" response. These two groups of children have different income distributions and should not be treated as a single group in imputation.

Income Imputation

Telephone surveys that have a single income question often use a single hot-deck donor pool or an overall regression model to impute missing values. Our general strategy for income imputation was to use regression imputation for Groups 1 to 9 and hot-deck imputation for Groups 10 to 16, where we already know family income within $10,000 or $5,000.

We examined two regression imputation approaches. In the first approach we developed a separate regression model for each of the nine groups. In the second approach we developed a single overall imputation model. Thus, we could compare imputation results for the more common approach of using an overall imputation model versus using separate models that take advantage of the partial income information. As a basis for developing the models we assembled the data as follows.

For Group 1 (no partial income information) we used all children in the 2000 NIS with a nonmissing value on the family income composite variable. For Groups 2 to 7 we used children with a family income in the specific interval (e.g., for Group 2 all children with a family income above $60,000). For Groups 8 and 9 we used all children with a "DK" or "Refusal" response to total family income, respectively, but a nonmissing value on the family income composite variable (i.e., they completed the cascading questions).

All regression models used the 2000 NIS data analysis weight, and the dependent variable was the log (base 10) of reported family income. Family incomes of $0 were recoded to $1. The maximum family income was truncated to $1,000,000 for the model development. For children that completed the income cascading questions, we used the midpoint of their income interval. For children in the top income category (greater than $75,000) we used the category median of $100,000.

The 34 predictor variables offered in the nine stepwise regression models included child characteristics, characteristics of the mother, family and household characteristics, and ecological variables related to the characteristics of the telephone exchange (i.e., area code/central office code) of the sample telephone number.

One goal for the regression imputation was to have a parsimonious set of models. Using the Schwartz criterion (Schwarz, 1978), we first determined a stopping point for each of the nine stepwise searches. Predictor variables that appeared in two or more of the nine models were included in the combined model. If a predictor appeared in only one model but had a t-statistic greater than 3.0, that variable was also included in the combined model. The final model included the 25 predictor variables listed in Table 2.

To assess the ability of the nine regression models to predict family income and to compare those results with the prediction from the overall (Group 1) model, we used the 1999 NIS sample. For each of the nine groups we divided the children in the 1999 NIS into two random halves. We used the first half to fit the model with the 25 predictor variables. We then used the second half as the validation sample and obtained predicted values of family income, which were compared with the reported values. We did the same for the overall (Group 1) model, which we applied to Groups 2 to 9. Table 3 shows the mean and median difference between the predicted income and reported income and also the interquartile range for the differences.

The results for the 1999 validation sample show that the separate models are able to predict family income more accurately than the overall model for Groups 2-5 and for Group 7. The results are mixed for Groups 6, 8 and 9. For Group 6 the separate model produced a larger mean difference, but the median difference and the interquartile range were smaller. For Group 8 the separate model produced a larger mean difference and interquartile range but a smaller median difference. For Group 9 the separate model had a larger median difference, but the mean difference and the interquartile range were smaller. Table 3 shows that the mean differences for the overall and separate models are substantial in most of the groups.

For Group 8, Figure 1 plots the difference from the separate model against reported income. The systematic pattern reveals a tendency for the model to overpredict, progressively, larger family incomes. A similar pattern was observed for the other groups.

The combined regression model developed for 2000 was applied to the children in the 2000 NIS in Groups 1 to 9 with missing values of the family income composite variable. For Groups 10 to 16 a weighted sequential hot-deck procedure was used to impute family income. The following three variables (the first three predictors to enter the overall Group 1 stepwise regression model) were used to form the imputation cells: WIC participation, race/ethnicity of the child, maternal education.

We compared the 2000 NIS weighted distribution of the family income composite variable, excluding the children with a missing value of the family income composite variable, with the income distribution that included their imputed values of family income. Table 4 indicates that the imputation results in a somewhat higher percentage of children with lower incomes: 31.8% versus 34.2% of family incomes are $20,000 or lower.

Summary

The income questions used in the NIS made it possible to develop separate regression imputation models for the partial income information groups. In general, those models provided more accurate imputed income values than the overall regression model. For the groups where the cascading questions placed family income in an income interval, we found that the separate regression models produced imputed values that were always within the target income interval. The overall model, on the other hand, produced some imputed values outside the target interval. We were also able to take advantage of the NIS questionnaire structure to create separate regression models for children with a "DK" or "Refused" response to the total family income question and no cascading information. Our combined regression imputation model included 11 telephone exchange characteristics.

References

Olson, Lorayn, Rodén, Ann-Sofi, Dennis, J. Michael, Cannarozzi, Francine, and Wright, Robert A. 1999. Alternative methods of obtaining family income in RDD surveys. *1999 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp. 940-945.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, 6, pp. 461-464.

Table 1. Partial Income Information Groups

Group Number	Group Definition	Percentage of Children with Missing Family Income Composite Variable (n=4,829)
1	No income information available	8.7
2	Income greater than $60,000	2.2
3	Income greater than $40,000	1.5
4	Income greater than $20,000	9.2
5	Income from $0 to $20,000	5.9
6	Income from $20,001 to $40,000	3.5
7	Income from $40,001 to $60,000	2.0
Subtotal (2-7)		24.2
8	Gave "DK" response to total family income question and no partial income cascading information is available	30.9
9	Gave "Refused" response to total family income question and no partial income cascading information is available	21.4
Subtotal (8-9)		52.3
10	Income from $40,001 - $50,000	1.2
11	Income from $30,001 - $40,000	1.6
12	Income from $20,001 - $30,000	1.5
13	Income from $15,001 - $20,000	1.5
14	Income from $10,001 - $15,000	1.3
15	Income from $10,001 - $20,000	4.2
16	Income from $0 to $10,000	3.5
Subtotal (10-16)		14.9

Table 2. Predictors in the Combined Regression Model

Child's Characteristics:

- Child ever received benefits under the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) benefits
- Child's race/ethnicity
- Household report of 4:3:1:3 up-to-date vaccination status
- Shot card used to report vaccinations

Mother's Characteristics:

- Mother's marital status
- Mother's education
- Mother's age

Family Characteristics:

- Geographic mobility status
- Residence in Metropolitan Statistical Area
- Number of persons in the household
- Number of children in the household
- Relationship of respondent to the child
- Household experienced an interruption in telephone service
- CATI system language queue assignment

Telephone Exchange Characteristics:

- Percent of households owner occupied
- Log of median home value
- Log of average rent
- Percent college graduate
- Median years of education
- Log of median household income
- Percent of households with income less than $10,000
- Percent of households with income from $15,000 to 24,999
- Percent of households with income from $35,000 to 49,999
- Percent of households with income from $50,000 to 74,999
- Percent of population that is Hispanic

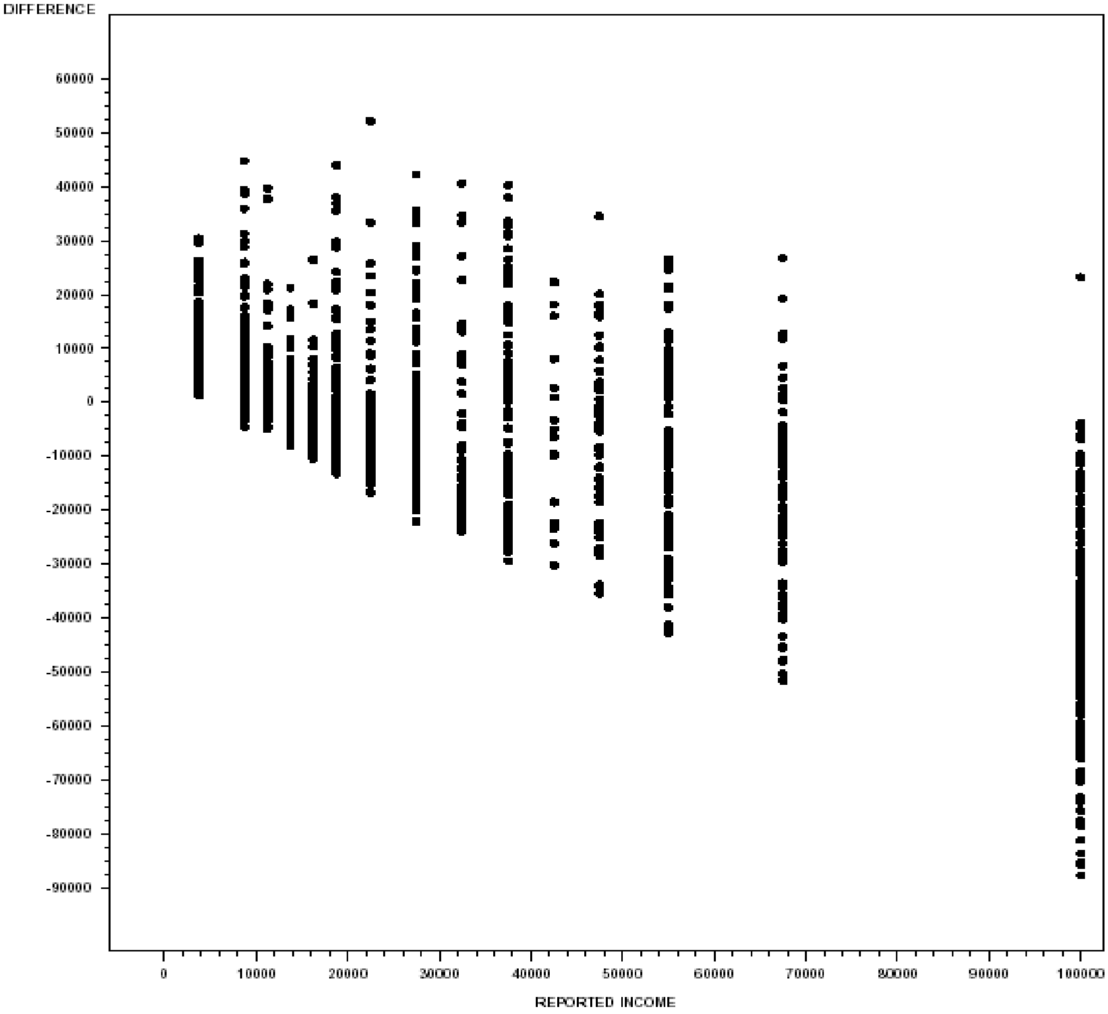
Table 3. Results for Validation Sample – Difference between Predicted and Reported Family Income (in dollars)

	Separate Models			Overall Model		
Group	Mean	Median	Interquartile range	Mean	Median	Interquartile range
1 (no information)	−8,275	−2,038	21,807	−8,275	−2,038	21,807
2 (>$60,000)	−8,596	5,926	29,519	−39,349	−25,935	35,760
3 (>$40,000)	−7,699	2,957	28,190	−22,787	−13,918	33,788
4 (>$20,000)	−5,602	1,535	24,910	−12,829	−7,614	26,409
5 ($0-20,000)	−2,785	−2,982	8,734	5,097	3,592	10,067
6 ($20,001-40,000)	−713	−528	8,701	460	−3,599	21,125
7 ($40,001-60,000)	−670	−45	9,095	−2,192	−685	27,099
8 (DK to total family income)	−3,904	−391	16,079	−1,287	894	15,407
9 (Refused total family income)	−4,886	−3,075	29,070	−5,289	−2,518	29,456

Table 4. Weighted Income Distribution Before versus After Imputation

Family Income	Excluding children with a missing income (%)	Including the imputed income of children with a missing income (%)
$0 - $7,500	6.9	6.8
$7,501 - $10,000	6.5	7.0
$10,001 - $12,500	3.4	4.6
$12,501 - $15,000	5.0	5.5
$15,001 - $17,500	2.9	3.3
$17,501 - $20,000	7.1	7.0
$20,001 - $25,000	7.5	7.3
$25,001 - $30,000	8.5	8.3
$30,001 - $35,000	5.4	5.4
$35,001 - $40,000	6.6	6.1
$40,001 - $45,000	3.9	3.9
$45,001 - $50,000	6.1	5.8
$50,001 - $60,000	8.1	7.9
$60,001 - $75,000	8.1	8.0
$75,001 or greater	14.2	13.1

Figure 1.(Reported - Actual) Income vs Reported Income -
Separate Model for Group 8 (DK to total family income)


DIFFERENCE
REPORTED INCOME