NBER WORKING PAPER SERIES

USING ADMINISTRATIVE DATA TO IMPUTE INCOME NON-RESPONSE IN HOUSEHOLD SURVEYS

V. Kerry Smith
Michael P. Welsh
Richard Carson
Stanley Presser

Working Paper 30420
http://www.nber.org/papers/w30420

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2022

The Data from the Deep Water Horizon Contingent Valuation Survey were developed with the support of the National Oceanic and Atmospheric Administration for a natural resource damage case. The case was settled and to the best of my knowledge the data are publicly available along with all the materials associated with their development. There was no support from this effort for this new research. The SOI data Internal Revenue Service are in the public domain and the website is listed in the paper. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Using Administrative Data to Impute Income Non-Response in Household Surveys
V. Kerry Smith, Michael P. Welsh, Richard Carson, and Stanley Presser
NBER Working Paper No. 30420
September 2022
JEL No. C0,C8

ABSTRACT

Income is simultaneously one of the most important variables used by economists and the variable most likely to be missing due to item non-response. While observations that are missing income responses are often dropped from analyses, such treatment is usually inappropriate. More appropriate solutions rely on imputation based on either covariates (e.g., age and education) measured in the survey or on spatial estimates (most often for zip codes) from the American Community Survey. We describe a new spatially-based alternative using publicly available Internal Revenue Service tax data that allows estimates of zip code's income distribution.

V. Kerry Smith
Department of Economics
W.P. Carey School of Business
P.O. Box 879801
Arizona State University
Tempe, AZ 85287-9801
and NBER
kerry.smith@asu.edu

Michael P. Welsh
1598 Venice Lane
Longmont, Colo 80503
welsh@mailbag.com

Richard Carson
Department of Economics
University of California, San Diego
ECON 323
9500 Gilman Dr. #0508
La Jolla, CA 92093-0508
rcarson@ucsd.edu

Stanley Presser
Department of Sociology and Joint Program
in Survey Methodology
1218 LeFrak Hall
Joint Program in Survey Methodology
College Park, MD 20742
spresser@socy.umd.edu

I. Introduction

Missing responses for income in surveys have long posed a problem because that variable is crucial to much of the analysis done by economists and other social scientists (Moore, et al. 2000). The simplest way of dealing with the problem is to drop the observations that are missing income. This practice is problematic unless the observations are missing completely at random, which is usually not true. Consequently, researchers often impute missing values, either by predicting income from covariates available in the survey, such as age and education, or by using spatially-based estimates such as median income from the U.S. Census Bureau's American Community Survey (ACS). Although generally superior to simply dropping cases that are missing income, these approaches also make assumptions that are unlikely to be true.

The error from imputing income has become larger during the past quarter century because the proportion of missing cases has grown dramatically during that time. Using data from the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP), for example, Meyer et al. [2015] estimated that item nonresponse (and therefore imputation) for income questions about six transfer programs increased 0.4 percentage points a year from 1991 to 2013, reaching over 40% for both Unemployment Insurance and Supplemental Security Income in the most recent year. Similarly, Bollinger et al. [2019] reported that imputation for total earnings in the CPS more than doubled from less than 10 percent in 1987 to almost 25 percent in 2015. Thus efforts to improve imputation of income have become increasingly important over time.

This note advances a new, spatially-based approach to imputation by demonstrating how publicly available data for zip codes from the Internal Revenue Service (IRS) Statistics of Income (SOI) program can be used to assign missing income. The SOI essentially has no item missing data and its reports should be of higher quality than those in surveys due to the potential penalties

for reporting inaccurately. On the other hand, the SOI has substantial unit missing data (because many households do not file a tax return), an issue we will return to after presenting our results.

Our approach has four ingredients. The first is a *common* spatial definition for the data in which income is imputed and the data used as the source of imputation information. In most instances, this will be the zip code, the lowest level of spatial aggregation for which the IRS releases summaries of adjusted gross income (AGI) from tax returns.[1] Zip code is often also the lowest spatial indicator in survey data, available from either the sample design or respondent reports.[2] Second, the IRS frequency distributions are used to estimate the parameters of the Dagum [1977] distribution for the reported AGI before taxes in each zip code for the year about which the survey respondents were asked to report their income. These parameters allow estimates of the income distributions of tax paying units at the zip code corresponding to their home communities.[3] Third, the incomes reported by survey respondents who answer the income questions are regressed on estimates for the parameters describing income distributions in the respondents' zip codes. We use the median income and the Gini coefficients derived from these distributions in our example.[4] Finally, predictions from the estimated model are used to impute income at the zip code level for those survey respondents who did *not* answer the income question.

We evaluate our imputation method using the national, in-person, survey of U.S. households conducted for the Deep Water Horizon (DWH) natural resource damage case (see Bishop et al. [2017]). We use a variation on a bootstrap strategy (Efron and Tibshirani [1993]) for each of three different sized, randomly selected, holdout samples (25, 50 and 100). Our analysis focuses on the sample of respondents who answered an open-ended question about their income before taxes. For the three different size holdout samples, between 72% and 81% of the 1,000

replications did not reject the null hypothesis that our estimated imputed income was an unbiased estimate of the reported income at the 95% significance level.

The next section provides the highlights of our example and summarizes our findings. The final section discusses the implications of our findings and suggests ways to extend what we have done in future work.

II. Methods, Data, and Results

We estimate a linear model relating income for surveyed households who answered the income question to the estimated median income and Gini coefficient derived from SOI frequency distributions for AGI at the zip code level. Separate estimates of the parameters of the Dagum distribution (and the associated estimates for median incomes and Gini coefficients) are developed for each zip code. These data are matched using the reported zip codes of survey respondents for the cases where SOI has sufficient records to support the public availability of these frequency distributions The median and Gini statistics are derived from the estimated parameters of the Dagum distributions fitted to the AGI frequencies.[5]

The DWH survey we used to gauge the performance of our proposal was administered in-person to a probability sample of households in the contiguous US that included at least one English-speaking adult between October 2013 and July 2014.[6] The income question asked about income in the year preceding the date of each interview. So respondents asked about income in 2012 were matched to SOI records for that year. Similarly, those asked about their 2013 income were matched to SOI records of the frequency distributions for AGI at the zip code level for that year.

The survey contained a sequence of three income questions.[7] The first was an open-ended question that asked respondents for their before-tax household income to which 68% (n=2,481)

provided an amount. Respondents who did not answer this first item were then asked to choose one of 10 intervals ranging from less than $10,000 to more than $200,000 and 23% (n=855) did so. The remaining respondents were then asked "Was it more than $35,000?" and, if so, "Was it more than $75,000?" Four percent (n=149) were coded this way, leaving 5% (n=171) who refused to provide any information about their household income.

Here we only use respondents who answered the first, open-ended income question to compare to the IRS estimates of the median income and Gini coefficients because the open item provides detail at the same level as the IRS data[8]. Three bootstrap simulations were conducted to evaluate our proposed imputation method. Each had 1,000 replications. They are distinguished by the size of the holdout samples (25, 50, and 100), with the records selected randomly without replacement. In each case the remaining sample was used to estimate how the respondent's reported income related to the Dagum estimates of median income ($\tilde{y}_i$) and the Gini coefficient (g_i) for the corresponding zip codes, based on the IRS SOI frequency distributions, as given in equation (1).

$$y_i = \alpha + \beta\tilde{y}_i + \gamma g_i + u_i \qquad (1)$$

i indexes the observation in the sample of DWH respondents who answered the direct income question and lived in zip codes that could be matched to the IRS SOI frequency distributions for AGI in the year for each person's reported income. The number of observations corresponds to this sample less the holdout sample. $\tilde{y}_i$ and g_i vary by zip code. The zip code is the most granular, publicly available, information on each respondent's location. As a result, in cases where some respondents come from the same zip code, the values for these variables will not change for them.

Predictions from equation (1) are used to impute the income for those records in each holdout sample. We treat each of these observations *as if* the respondent did not report income. Of

course, in these cases we know what the respondents actually reported. Thus, we can compare the imputations with the actual income responses. Equation (2) provides the form of our test. Reported income (y_i) is regressed on the predicted income ($\hat{y}_i$) from equation (1) using the values for the median income and Gini coefficients derived from the Dagum parameters based on the IRS frequency distributions. The joint hypotheses that *c*=0 and *d*=1 provides the basis for our evaluation of the imputation procedure.

$$y_i = c + d\hat{y}_i + \varepsilon_i \,. \qquad (2)$$

We assume the errors (u_i and ε_i) in each model are independent. The two models (equations (1) and (2)) might be interpreted as implying that the income responses (y_i) and income predictions ($\hat{y}_i$) are random variables. While income is assumed to be a random variable, predicted income should be treated as akin to an instrument for potential imputed values of income such as the values that might be derived from a hot deck.[9] Thus, the textbook argument holding that if we assume there are errors in an independent variable, then those errors lead to biases that "reduce" the expected value for the estimates of *d* do not apply.[10]

Failure to reject the composite hypothesis implies we cannot distinguish the predictions from the survey responses. We used a p-value of 0.05 in all our tests. Table 1 provides examples of the estimates for equation (1) and equation (2) using one randomly selected holdout sample. Figures 1a through 1c provide the kernel density estimates for the estimated values of the coefficient associated with predicted income (d) using the 1,000 replications for each of the three different sized holdout samples. All of the figures indicate the estimates of *d* are approximately centered at one. Our tests show 81% fail to reject the joint null hypothesis using a holdout sample of 25, 74.5% with a sample of 50, and 72% with a sample of 100.

The IRS SOI frequency distributions are publicly available and offer one strategy for remediating the bias from unit nonresponse and potentially item nonresponse to income questions. Sabelhaus et al. [2015] found unit nonresponse was more likely in the highest income households with the top 5 percentiles averaging 66 percent and the top 1 percent by AGI having a 65 percent response rate. Of course, this does not address the issues raised by those who do not file returns. A possible solution to this problem would combine the results from the SOI with estimates from the ACS for the non-filers (those missing in the SOI) in the relevant zip codes – and then make the imputation based on the combined information. Given that one can ask whether the respondent's household filed a return (which is not likely to suffer from substantial missing data), this combined strategy would use the SOI based estimates to impute for those who did file and the ACS zip code estimate (adjusted using a model predicting filing status) for those who report their household did not file a return. Such a combined approach has not to our knowledge been considered.

III. Implications and Extensions

Researchers often do not have access to the information that would be needed to assemble a detailed spatial map of the survey respondents. This makes using hot-deck type procedures that exploit fine (e.g., street level) spatial homogeneity difficult, if not impossible, to implement. Our proposal calls for the use a different source of information on the lowest level spatial indicator included in many surveys, the zip code, for which high quality income data are released by the IRS.

One implication of using the IRS AGI income definition as the imputation source is that the survey's income question should mirror a simplified version of its definition. We chose to use the DWH survey income question for the illustration, in part, because its income question was

worded to that end. The translation from the income variable in the survey to any specific definition of income like AGI can be important to model.

The two current common methods for income imputation are: statistical models based on those survey respondents who report incomes and the hot deck method. While we don't make any specific claim about the performance of our proposed approach relative to these two approaches, some issues can be identified in comparing them to our method. For the first one might ask whether, for example, the income levels of a single, college educated, thirty-year old in a rural Mississippi zip code versus a person with the same demographic characteristics in a downtown San Diego zip code are likely to be similar. The answer seems very likely to be no. Less clear though is what would happen if state level geography is used. One might well find that the demographic predictors explain more of the variation in income. Likewise, as a survey gathered more information, such as marital status and (if married) whether the spouse works for pay, and as more effort goes into building a better predictive model of income, the better the relative performance of the demographic based predictor.

A somewhat different perspective on why the spatial approach might be preferred has to do with the well-known income stratification that impacts locational choice decisions (Kuminoff et al. [2013]). The combination of results from Bee et al. [2015] and Sabelhaus et al. [2015] suggest that the biases in income measures may be especially important where sorting and associated income stratification impact the realized income distributions. In such cases, the estimated role for household income in explaining the outcomes that are the focus of any specific analysis may well be affected when imputation methods fail to account for the role of behavioral influences that could lead to the location specific differences in income distributions. That is, Tiebout sorting, along

with the assumption that preferences satisfy the single crossing condition, imply differences in local public goods across communities can lead to income stratification.

A different sort of comparison is between using the ACS's median zip code income estimate and using the IRS mean AGI estimate for the zip code as the imputed value. Both are spatial estimates. They vary in two ways. AGI and Census income definitions, while similar are not quite the same, and one is a median estimate while the other is a mean estimate.[11] The latter divergence is not a necessary element of our procedure. Because it is the parameters of the Dagum distribution that are being estimated the mean, or any income quantile, including the median can be used.

Indeed, given the Dagum distribution parameter estimates it is possible to draw repeatedly from it in a manner akin to multiple imputation. At a simple level, then, the missing income variable is replicated k times with the only difference being that each of these observations has the same covariate values except for income, each of which uses a different draw, with each of these k replicants receiving a weight of 1/k rather than 1.[12] Uncertainty over the actual value of income is now incorporated into the model's parameter estimates. In some instances, the Dagum distribution's parameter estimates, often cast in Gini coefficient form, may be sufficient to characterize this uncertainty.

The more substantive differences between the ACS and IRS income measures as the imputation source in a survey is likely not definitional differences since many survey respondents do not think about a comprehensive definition of income.[13] The use of the IRS tax return data eliminates the twin problems in the ACS caused because the income questions have high non-response rates and reporting errors that are known to be widespread and often large. Those

advantages are cast against tax returns (including those filed to claim tax credits) while covering a large majority of those taking surveys, have some divergences in coverage.[14]

Statistical precision might be improved by knowing how heterogenous income is within a zip code. While the IRS is unlikely to ever publish Census block-group level income data, knowing the distribution of Census Bureau block groups' parameter estimates for the Dagum distribution relative to that for the zip code might be a sufficient statistic for many purposes. The underlying intuition for this proposal stems from the assumption that the income distribution within a zip code is more homogenous than across zip codes. Access to the block-group level variation would allow this hypothesis to be tested.

Finally, more accurate imputed income values are most likely to come by replacing our zip code level data with zip code level data that are split into two groups whose income distributions are distinct. For example, separate estimates of the parameters of the Dagum distribution for single individual households versus two-or-more person households.[15] From a survey perspective the task would be to ask the conditioning question in a manner that matches the methods the IRS might use to form the partitioning into two groups. While the income on IRS tax returns can be decomposed in a number of ways, such as considering the presence or absence of stock dividend income, the relevant possibilities are those where there are overlaps in the information available on standard surveys and on the IRS tax forms.

Table 1: Simple Imputation Model and Out-of-Sample Test of Imputation Performance

Independent Variables, Fit, and Sample Size	Imputation Model	Test Model for Evaluating Imputation Performance
Median Income Estimate by zip code	1.20 (15.56)	…
Gini Coefficient Estimate by zip code	1.51×10^5 (5.59)	…
Predicted Income for Missing Values	…	1.10 (2.17)
Intercept	-80.32×10^3 (-5.61)	1.01×10^4 (0.25)
R^2	0.144	0.170
Number of Observations	1850	25

a. Numbers in parentheses below estimated coefficients are the t-statistics for the null hypothesis of no association constructed using robust standard errors.

Figure 1. Kernel Density Estimates for Holdout Samples of Sizes 25, 50 and 100

NOTE: Figures 1a, 1b, 1c will be one three panel figure

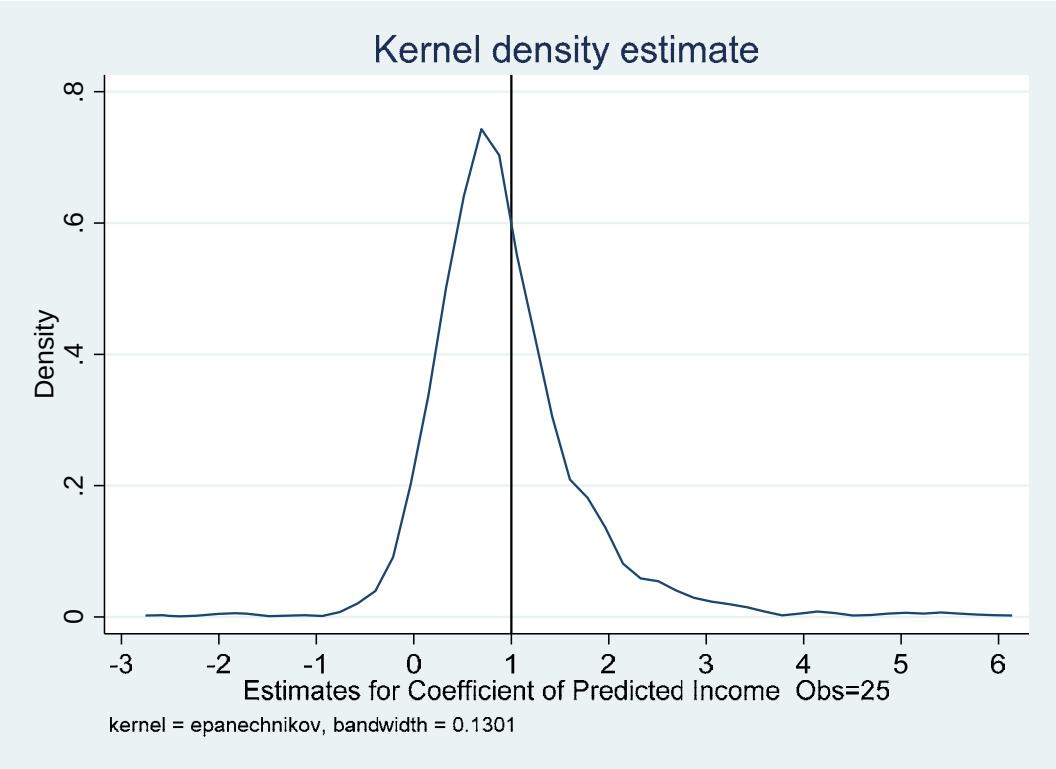

Figure 1a : Kernel Density for Estimates of d with Holdout sample =25

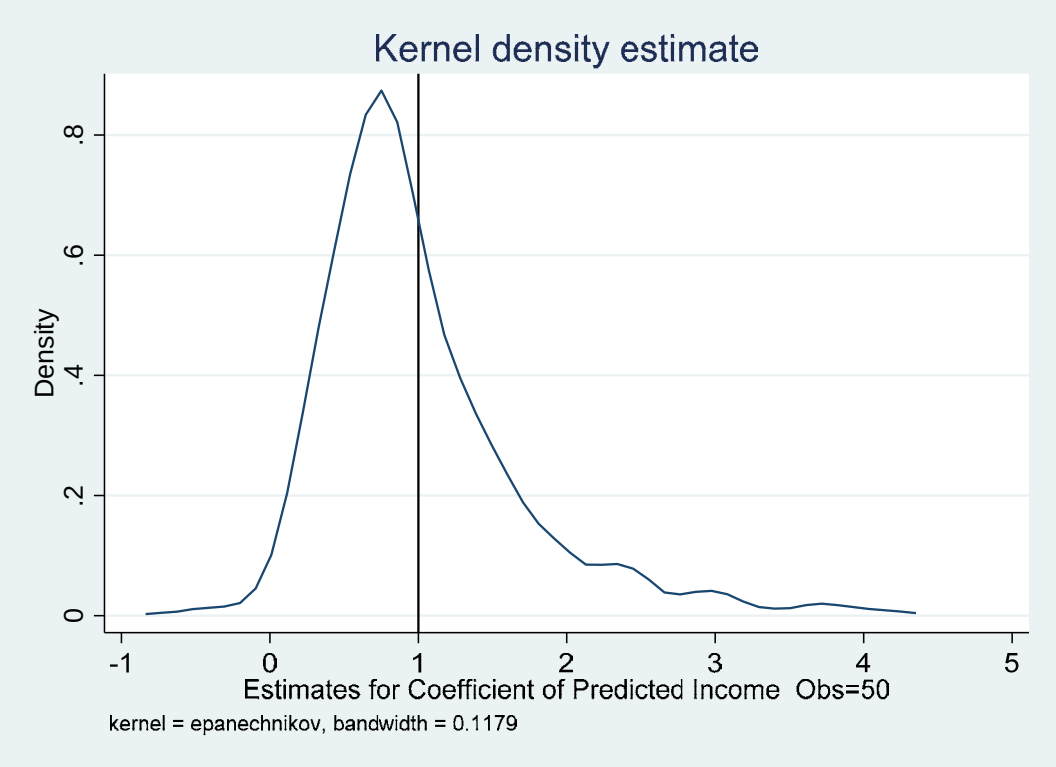

Figure 1b: Kernel Density for Estimates of d with Holdout sample of 50 observations

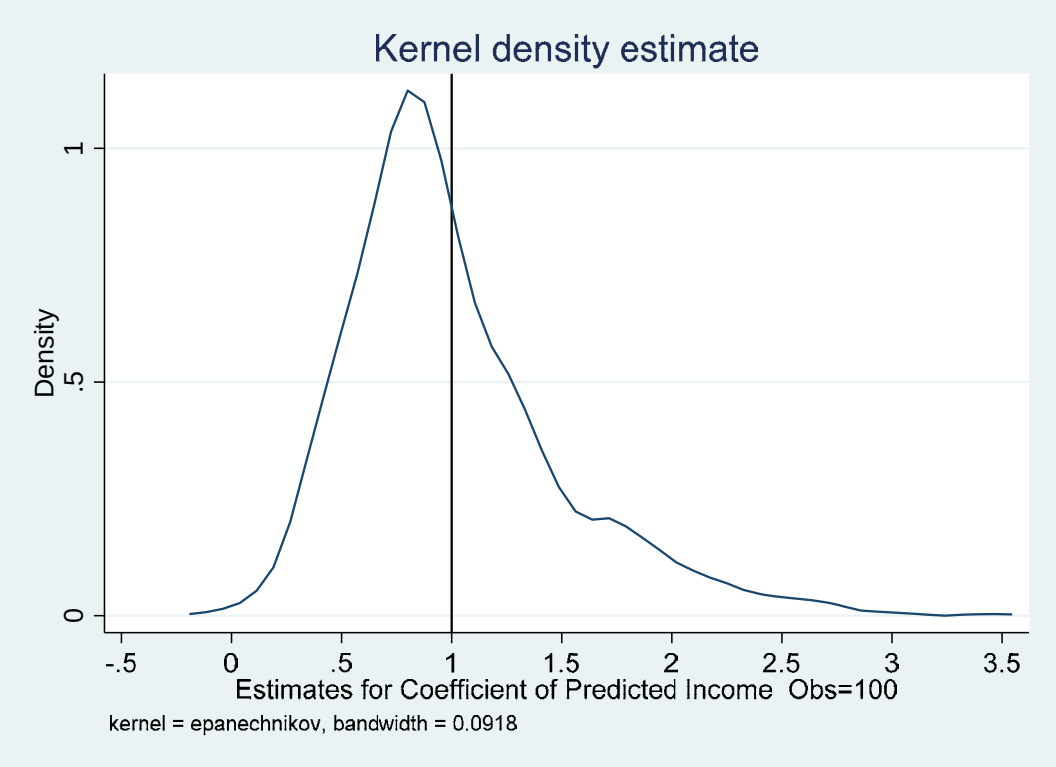

Figure 1c: Kernel Density for Estimates of d with Holdout sample of 100 observations

Appendix A: Wording of the Three Income Questions

[IF HHFAMSIZE = 1] I now need to know what your total income was for all of [IF DATE IS DEC 31, 2013 OR EARLIER: 2012; IF DATE IS JAN 1, 2014 OR LATER: 2013].
This includes income from jobs, pensions, social security, interest, dividends, capital gains, profits from businesses, unemployment payments, and **all** other money you received.

[IF HHFAMSIZE>1] I now need to know what your total family income was for all of [IF DATE IS DEC 31, 2013 OR EARLIER: 2012; IF DATE IS JAN 1, 2014 OR LATER: 2013].
This includes income from jobs, pensions, social security, interest, dividends, capital gains, profits from businesses, unemployment payments, and **all** other money you received.
Your total family income includes your own income plus the incomes of all family members who live with you.

[IF HHFAMSIZE=1] Adding up the income from all these sources and all other sources, what was your total income for all of [IF DATE IS DEC 31, 2013 OR EARLIER: 2012; IF DATE IS JAN 1, 2014 OR LATER: 2013], before taxes?

[IF HHFAMSIZE>1] Adding up the income from all these sources and all other sources, what was the total income for you and your family living with you for all of [IF DATE IS DEC 31, 2013 OR EARLIER: 2012; IF DATE IS JAN 1, 2014 OR LATER: 2013], before taxes?

IF RESPONDENT ANSWERS DON'T KNOW, GIVE A VERY LONG PAUSE AND THEN SAY: It would be a big help to us if you would be willing to give me your best estimate in answering the question, even if you're not completely sure. **[REPEAT QUESTION].**

Q21v1. PROGRAM VALIDATION SCREEN: "I typed $"**XX**" dollars per year. ($"XX" should be bolded and large type). Is this what you said?"

YES 1
NO 2
(vol) REFUSED-97
(vol) DON'T KNOW-98
IF Q21v1=NO: RE-ENTER AMOUNT
IF Q21v1=YES AND Q21 AMOUNT EQUAL 0 THEN: Q21v2. Just to check, you **[IF HHFAMSIZE>1: and your family living with you]** received no money from **any** source in [IF DATE IS DEC 31, 2013 OR EARLIER: 2012; IF DATE IS JAN 1, 2014 OR LATER: 2013], is that correct?

YES 1
NO 2
(vol) REFUSED-97
(vol) DON'T KNOW-98
[IF R PROVIDES AMOUNT, SKIP TO NEXT SECTION]
IF R REFUSES OR DK INCOME: ASK INCOME CATEGORY QUESTION Q22.

Q22 Could you please just tell me the letter on this screen that best matches **[(IF HHFAMSIZE=1)**: your total income for all of [IF DATE IS DEC 31, 2013 OR EARLIER: 2012; IF DATE IS JAN 1, 2014 OR LATER: 2013]/ **(IF HHFAMSIZE>1)**: the total income for all of [IF DATE IS DEC 31, 2013 OR EARLIER: 2012; IF DATE IS JAN 1, 2014 OR LATER: 2013] for you and your family living with you**]**?

A. Less than $10,000 **1**
B.$10,000 to $14,999 **2**
C.$15,000 to $24,999 **3**
D.$25,000 to $34,999 **4**
E.$35,000 to $49,999 **5**
F.$50,000 to $74,999 **6**

G.$75,000 to $99,999 **7**
H.$100,000 to $149,999**8**
I.$150,000 to $199,999 **9**
J.$200,000 or more **10**
(vol) REFUSED-97
(vol) DON'T KNOW-98
IF R REFUSES OR DK INCOME ASK INCOME CATEGORY QUESTION Q23A.

Q23A Could you please just tell me whether the total income for all of [IF DATE IS DEC 31, 2013 OR EARLIER: 2012; IF DATE IS JAN 1, 2014 OR LATER: 2013], for you **[(IF HHFAMSIZE>1)**: and your family living with you], was more than $35,000?
YES 1 Ask Q23B
NO 2 SKIP TO NEXT SECTION
(vol) REFUSED-97 SKIP TO NEXT SECTION
(vol) DON'TKNOW-98 SKIP TO NEXT SECTION

Q23B Was it more than $75,000?
YES 1 SKIP TO NEXT SECTION
NO 2 SKIP TO NEXT SECTION
(vol) REFUSED-97 SKIP TO NEXT SECTION
(vol) DON'T KNOW-98 SKIP TO NEXT SECTION

REFERENCES

Andridge, Rebecca R., and Rodrick J.A. Little 2010, "A Review of Hot Deck Imputation for Survey Non-Response," *International Statistical Review*, 78(1), pp. 40-64.

Bee, Adam, Bruce D. Meyer, and James X. Sullivan, 2015, "The Validity of Consumption Data: Are the Consumer Expenditure Interview and Diary Surveys Informative?" in *Improving the Measurement of Consumer Expenditures* edited by Christopher D. Carroll, Thomas F. Crossley, and John Sabelhaus (Chicago: University of Chicago Press for NBER), pp.204-240

Bishop, Richard C., Kevin J. Boyle, Richard T. Carson, David Chapman, W. Michael Hanemann, Barbara Kanninen, Raymond J. Kopp, Jon A. Krosnick, John List, Norman Meade, Robert Paterson, Stanley Presser, V. Kerry Smith, Roger Tourangeau, Michael Welsh, Jeffrey M. Wooldridge, Matthew De Bell, Colleen Donovan, Matthew Konopka, and Nora Scherer, 2017, "Putting a Value on Injuries to Natural Assets: the BP Oil Spill," *Science*, April 21, pp. 253-254.

Bollinger, Christopher R., Barry Hirsch, Charles M. Hokayem, and James P. Ziliak, 2019, "Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years after Lillard, Smith, and Welch," *Journal of Political Economy*, vol 127, pp. 2143-2185.

Dagum, C., 1977, "A New Model of Personal Income Distribution: Specification and Estimation," *Economie Appliquee*, vol 30, pp. 413-437.

Elwell, James, Kevin Corinth, and Richard V. Burkhauser, 2021, "Income Growth and Its Distribution from Eisenhower to Obama: The Growing Importance of In-Kind Transfers (1959-2016)," in Diana Furchtgott-Roth (editor), *United States Income, Wealth, Consumption, and Inequality* (New York: Oxford University Press) pp. 90-124.

Efron, Bradley and Robert J. Tibshirani, 1993, *An Introduction to the Bootstrap* (New York: Chapman Hall).

Greene, William H., 2012, *Econometric Methods*, seventh edition, (New York: Pearson).

Henry, Eric L., and Charles D. Day, 2005, "A Comparison of Income Concepts: IRS Statistics of Income, Census Current Population Survey, and BLS Consumer Expenditure Survey," *IRS Methodology Report Series, Special Studies in Federal Tax Statistics*, pp. 149-157.

Herriot, Roger A., 1977, "Collecting Income Data on Sample Surveys: Evidence from Split-Panel Studies" *Journal of Marketing Research, Special Issue: Recent Developments in Survey Research,* Vol.14 (3) pp. 322-329.

Kleiber, Christian, 1996, "Dagum vs. Singh-Maddala Income Distributions," *Economics Letters*, Vol.53, pp. 265-268.

Kleiber, Christian, 2007, "A Guide to the Dagum Distributions," WWZ Working Paper, No. 23/07, University of Basel, Center of Business and Economics (WWZ),Basel.

Kuminoff, Nicolai, V. Kerry Smith, and Christopher Timmins, 2013, "The New Economics of Equilibrium Sorting and Policy Evaluation Using Housing Markets," J*ournal of Economic Literature*, Vol. 51, December, pp. 1-57.

Lukasiewicz, Piotr, Krzysztof. Karpio, and Arkadiusz Orłowski, 2012, "The Models of Personal Incomes in USA," *Acta Physica Polonica A*, Vol. 121, pp. B82-B85.

Meyer, Bruce D., Wallace K.C. Mok, and James X. Sullivan, 2015, "Household Surveys in Crisis," *Journal of Economic Perspectives*, Vol. 29, Fall, pp.1-29.

Moore, Jeffrey C., Linda L. Stinson, and Edward J. Welniak, 2000, "Income Measurement Error in Surveys: A Review," *Journal of Official Statistics*, 16(4), pp. 331-362.

Sabelhaus, John, David Johnson, Stephen Ash, David Swanson, Thesia I. Garner,John Greenlees, and Steve Henderson, 2015, "Is the Consumer Expenditure Survey Representative by Income?" ?" in *Improving the Measurement of Consumer Expenditures* edited by Christopher D. Carroll, Thomas F. Crossley, and John Sabelhaus (Chicago: University of Chicago Press for NBER), pp.241-262.

Smith, V. Kerry, 2021, "A Note on Income Inequality at the State Level," *Economics Letters*, Vol. 203, pp. 1-3.

Steinberg, Dan, Richard T. Carson and Leo Breiman, 1998, "Broad Scale Missing Value Imputation with Iterative Binary Partitioning," in David W. Scott, ed., *Computer Science and Statistics: Proceedings of the 29th Symposium on the Interface* (Fairfax Station, VA: Interface Foundation of North America).

Tadikamalla, Pandu. R. 1980, "A Look at the Burr and Related Distributions," *International Statistical Review*, vol. 48(3), pp. 337-344.

---

[1]Zip code level data are available at: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi. The IRS definition for AGI is similar to Elwell et al.'s [2021: 95] definition of a tax paying unit's market income as including gross income from wages and salaries, farm income, self-employment and business income, income from pensions, dividends, interest, rent, and alimony. According to the IRS (https://www.irs.gov/), AGI includes wages, dividends, capital gains, business income, and retirement distributions. Adjustments to income include items such as some employment-related expenses, student loan interest, transfers related to alimony on which taxes have already been withheld, and tax-deferred contributions to a retirement account. Thus AGI will never exceed a person's gross income, and for the vast majority of households, will be reasonably close to what economists hypothesize household income should measure.

[2] Of course, in order to conduct in person interviews information on respondents' addresses must be available. These details are omitted in public data sets to assure confidentiality of survey respondents.

[3]The Dagum distribution has been found to provide a flexible fit for grouped (interval-censored) data on income distributions (see Kleiber [1996]). The Dagum distribution function is defined by:

$$F(y) = (\ 1 + ({}^{b}/_{y})^{a})^{-p}\ ,$$

with *y* restricted to be greater than 0 and the three parameters, *a*, *b*, and *p* restricted to be non-negative. Conceptually, other distributions that have been used to model income distributions could be used. The Dagum distribution from a pure curve fitting perspective has been shown to be among the best of these (see Lukasiewicz, Karpio and Orłowski [2012] and its parameters have convenient interpretations.

[4] The median for a random variable that follows a Dagum distribution is given by:

$$\text{median}\ = b((\frac{1}{2})^{-\frac{1}{p}} - \ 1)^{\frac{1}{a}}\ \ .$$

The Gini coefficient is:

$$g = \frac{\Gamma(p)\Gamma(2p + {}^{1}/_{a})}{\Gamma(2\text{p})\Gamma(\text{p} + {}^{1}/_{a})}\ \ - 1\ ,$$

where $\Gamma(p)$ is the Gamma function defined as:

$$\Gamma(\theta) = \int_{0}^{\infty} exp(u)\, u^{\theta-1} du\ .$$

[5] The Stata command dagfit contributed by Austin Nichols uses maximum likelihood to fit a three parameter Dagum distribution to grouped data. This distribution has been referred to by a number of different names in the statistics literature including the Burr Type 3, beta-k, and 3-parameter Kappa distribution (Tadikamalla 1980). See http://fmwww.bc.edu/RePEc/bocode/d/dagfit.hlp

[6] 3,656 people completed the survey for a response rate (AAPOR-RR3) of 48% (Bishop et al., 2017; p. 254). The data and documentation are available in the supplemental material for that paper.

[7] Appendix A shows the text for the three questions.
[8] The completed DWH sample has 3,656 cases. 2,481 of which answered the open ended income question. 854 (87.6%) of those who answered the open-ended income question and were interviewed in 2013 could be matched to SOI records at the zip code level. A smaller percentage of those interviewed in 2014 were matched: 1,028 (68.3%).

[9] The hot deck procedure's name stems from its original implementation using a key punch card reader. As the deck of cards went through the machine, the reader retained the value for a variable of the last card in case the current card's observation was missing. The original intuition was that if interviewing was done in clusters, then there would be a strong first-order spatial correlation in income. Variants of this procedure, including those that condition on a set of indicator variables, are now widely used, including by the U.S. Census Bureau's Current Population Survey [Aldridge and Little, 2010]. Steinberg et al. [1998] show how CART trees could be used to form optimal partitions for imputing missing values in the decennial U.S. Census.

[10] This interpretation stems from Greene's [2012] discussion (pp. 356-357) of using instrumental variables as a response to errors in variables problems. However, in our case we do not estimate the variance of ε_i by replacing $\hat{y}_i$ with y_i, as Greene emphasizes in his discussion of the uses of instrumental variable estimators for structural equations (pp. 366-367). Our objective here is imputation, so the estimated variability of ε_i should rely on what is used in the imputation. Replacing the instrument ($\hat{y}_i$) with y_i in computing the error variance for our holdout sample analyses would make it more difficult to reject the composite hypothesis, based on the kernel distributions for our estimates of d. If we were using the same logic to evaluate a hot deck imputation, then his argument would apply.
[11] A detailed comparison of similarities and differences in the income definition used by the IRS SOI, Census Bureau's Current Population Survey, and the Bureau of Labor Statistics Current Expenditure Survey is provided by Henry and Day and [2005].

[12] The Dagum parameter estimates for each zip code can also be used to estimate the probability that a particular observed income value came from a particular part of the zip code's income distribution (e.g., 5^{th} to 95^{th} percentiles). This logic underlies Smith's [2021] test for differences in the degree of income inequality over time at the state level using the IRS SOI data to estimate Dagum distributions for each state and year.
[13] Indeed, language in U.S. Census Bureau surveys (and our language in the DWH survey income question contained in Appendix A can be seen as reminding respondents of income sources that are often not reported like social security where it is common for some respondents to think that all they are doing is getting money out that they paid in earlier.

[14] This divergence is smaller than some might think at first, because in addition to those paying income taxes, tax returns are also filed by those qualifying for various government tax credits such as the earned income tax credit, health care subsidies, and renewable energy rebates. Missing in both the IRS data and almost all surveys are institutionalized populations like those in prison as well as the homeless and undocumented immigrants. The IRS SOI data is not reported for zip codes where less than 100 returns were filed, where the zip code is for a single building/complex or for non-residential zip codes.
[15] A national or state-level correction factor could be developed from occasional IRS releases of random samples of individual tax returns, which do not include a zip code level spatial identifier.