

NATIONAL HEALTH AND AGING TRENDS STUDY (NHATS)
Round 5 Income Imputation

December 1, 2016

Suggested Citation: DeMatteis Jill M., Freedman Vicki A., and Kasper, Judith D. 2016. National Health and Aging Trends Study Round 5 Income Imputation. NHATS Technical Paper #15. Johns Hopkins University School of Public Health. Available at www.NHATS.org. This technical paper was prepared with funding from the National Institute on Aging (U01AG032947).

Overview

In preparing survey data files for analysis, imputation is often used to address item nonresponse, particularly when complex multi-variate recodes are required that are built up from a collection of more detailed questions (Marker, Judkins, and Winglee, 2001). Rounds 1, 3, and 5 of the National Health and Aging Trends Survey (NHATS) include imputed values for total income. Both a continuous measure and a bracket value are provided, with separate bracket values for single respondents and those who are married or are living with a partner. We used a cyclical n-partition hot deck (see Judkins 1997) to generate five imputations of each measure. This technical paper provides details on the imputation strategy.

Income Sources Collected in NHATS

Round 1 of NHATS collected information on sources of income (yes/no) and amounts for each source. Round 3 also collected information on sources of income (yes/no) but not amounts for each source. Like Round 1, Round 5 included collection of information on both sources of income and amounts for each source. In all three rounds, respondents with a spouse/partner were given the option of reporting amounts for themselves either together or separately from their spouse/partner. Table 1 shows the income sources included in NHATS.

Table 1. Summary of Income Sources Collected in NHATS Round 3

Sources of Income	Time frame
Social Security	Last Month
Supplemental Security Income	Last Month
Veteran's Administration	Last Month
Pension plan	Last Month
Earned Income	Last Month/Last Paycheck
Interest/dividend income from any: mutual funds/stocks, bonds, bank accounts, or CDs ¹	Last Year
Retirement account withdrawals ¹	Last month/Last year
Total income from all sources	Last Year

¹Asked after questions about the existence of relevant asset

After reporting about the receipt of income from each individual source, respondents are asked to report total income from all sources. Respondents who report don't know or who refuse are offered a set of five bracketed ranges. For 2015, we developed ranges by combining the 2010 and 2013 Survey of Consumer Finance samples to create weighted distributions for the population ages 65 and older. Separate ranges were provided for individually reported (respondent or spouse/partner amounts separately) and for jointly reported (respondent and spouse/partner together) amounts.

Extent of Missing Data for Total Income

In 2015, fifty-six percent of the sample provided a total income amount and an additional 19% reported a bracketed value instead. Thus, a bracketed value could be created from reports for 75% of the sample and required imputation for 25%. An exact value was then imputed for 44% of the sample (19% within a reported bracketed value and 25% within an imputed bracketed value).

Imputation Methodology

Westat's AutoImpute software was used to impute five values of the total income items. AutoImpute uses a cyclical n-partition hot deck (an approach analogous to the Gibbs sampler but using the hot deck

to generate the imputations). (See Judkins 1997; Judkins et al. 2007; Judkins, Piesse, and Krenzke 2008; Krenzke and Judkins 2008.) This software is designed to facilitate preservation of multivariate distributions while also ensuring that imputations maintain skip patterns and adhere to constraints. In this application an example of a constraint is ensuring imputations for specific amounts fall within reported (or imputed) bracket ranges.

The cyclical n-partition hot deck procedure initially imputes all target variables (i.e., items requiring imputation) using a simple hot deck that uses specified auxiliary variables and skip controllers. Using the initial imputed variables, a model is fit for each target variable using simple forward stepwise regression selection. The predicted values of the target variable from the final model are used to generate imputed values by using predictive mean matching for ordinal (ordered categorical) target variables and clustering for unordered categorical target variables. Predictive mean matching uses a hot deck with the skip controllers as hard boundaries and the predicted values from the final model as soft boundaries. For unordered categorical target variables, a k-means clustering algorithm is used on the vector of predicted values for each level, and then a hot deck is used to impute the target variable with the skip controllers as hard boundaries and the cluster membership indicators as soft boundaries. For more details on the procedure see Judkins et al. (2007).

Variables Used in Imputation

Three classes of variables were used in the imputation (see Appendix tables):

1. Source variables that indicate (yes/no) whether the respondent (and his/her spouse/partner, if applicable) has the particular source of income (referred to below as “fencepost” variables);
2. Source amount variables that indicate the amount of income from the particular source; and
3. Auxiliary variables that included respondent characteristics (e.g., age, race/ethnicity, gender, educational attainment, home ownership (in Round 5), veteran’s status, labor force status (in Round 5), spouse/partner’s labor force status (in Round 5), interviewer observations about the home condition) and skip pattern controllers.

In order to preserve the joint distribution of the full set of income variables, all missing fencepost, source amount, auxiliary, and total income variables were imputed. For respondents with a spouse/partner, separate respondent and spouse/partner variables as well as the combined variable were imputed for each source.

When imputing total income variables, the Round 5 source variables and source amounts were used, along with the auxiliary variables. Total income value was also constrained to fall within the reported/imputed bracket amount.

Income Imputation Variables in the SP File

The following imputed variables are included on the SP data file:

Variable name	Label	Description
<i>2014 Total Income Value</i>		
ia5totinc	R5 IA50 TOTAL INCOME	Actual
ia5toincimf	R5 F IMPUTED TOTAL INC FLG	Flag indicating imputation
ia5toincim1-5	R5 IA50 IMPUTED TOTAL INC1-INC5	Imputed values 1-5
<i>2014 Total Income Range – Respondents who have spouse/partner</i>		
ia5toincesjt	R5 IA51A JOINT EST TOT INCOME	Actual
ia5eincimjf	R5 F IMPTD JOINT EST TOT INC FLG	Flag indicating imputation
ia5eincimj1-5	R5 IA51A IMP EST JOINT TOT INC1-INC5	Imputed values 1-5
<i>2014 Total Income Range – Respondents who are single</i>		
ia5toincessg	R5 IA51B SINGLE EST TOT INC	Actual
ia5eincimsf	R5 F IMPUTED SGL EST TOT INC FLG	Flag indicating imputation
ia5eincims1-5	R5 IA51B IMP EST SP SGL TOT INC1-INC5	Imputed values 1-5

Using the Five Versions of the Imputed Variable in Analysis

For each of the three total income variables that was imputed, five sets of imputed variables were generated. For item respondents, these five sets contain copies of the reported data. For item nonrespondents, the five sets contain five independently generated imputed values. These five sets of imputed variables are provided to enable data users to use multiple imputation variance estimators and analysis techniques (see, for example, Rubin 1996) to account for the effects of item nonresponse and imputation error in variance estimates for analyses that use these income variables.

References

- Judkins, D. (1997). Imputing for Swiss cheese patterns of missing data. *Proceedings of Statistics Canada Symposium '97*, 143-148.
- Judkins, D., Krenzke, T., Piesse, A., Fan, Z., and Haung, W.C. (2007). Preservation of skip patterns and covariate structure through semi-parametric whole questionnaire imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3211-3218.
- Judkins, D., Piesse, A., and Krenzke, T. (2008). Multiple semi-parametric imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.
- Krenzke, T. and Judkins, D. (2008). Filling in blanks: Some guesses are better than others – Illustrating the impact of covariate selection when imputing complex survey items. *Chance*, 21(3), 7-13.
- Marker, D. A., Judkins, D. R., and Winglee, M. (2001). Large-scale imputation for complex surveys. In *Survey Nonresponse*, Eds. R. M. Groves, D. A. Dillman, E. L. Eltinge, and R. J. A. Little. New York: John Wiley & Sons.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.

Appendix. Lists of Variables Used in NHATS Round 5 Income Imputation

Table 1. Round 5 Source (“Fencepost”) Variables

#	Variable name	Label	% missing
1	ia5recsspa1	R5 IA1 SP REC SOCIAL SECURITY	4.7
2	ia5recsspa2	R5 IA1 SPOUSE PART REC SOC SEC	4.7
3	ia5recsspa3	R5 IA1 NO SOC SECURTY PAYMNT REC	4.7
4	ia5recssils1	R5 IA4 SP RECEIVD SSI LAST MONTH	5.8
5	ia5recssils2	R5 IA4 SPOUSE PRT REC SSI LST MO	5.8
6	ia5recssils3	R5 IA4 NO SSI RECEIVD LAST MONTH	5.8
7	ia5rvapayls1	R5 IA5 SP REC PAY FRM VA LAST MO	4.6
8	ia5rvapayls2	R5 IA5 SPOUS PA REC VA PAY LSTMO	4.6
9	ia5rvapayls3	R5 IA5 NO VA PAY REC LAST MONTH	4.6
10	ia5penjobou1	R5 IA6 SP HAS PENSION PLAN	5.7
11	ia5penjobou2	R5 IA6 SPOUSE HAS PENSION PLAN	5.7
12	ia5penjobou3	R5 IA6 NO PENSION PLAN	5.7
13	ia5iraothac1	R5 IA7 SP HAS IRA OTH RETIRE ACC	7.6
14	ia5iraothac2	R5 IA7 SPOUSE HAS IRA OTHR ACC	7.6
15	ia5iraothac3	R5 IA7 NO IRA OTHR RETIRE ACCT	7.6
16	ia5mutfdstk1	R5 IA8 SP OWNS MUTUAL FUND STOCK	7.1
17	ia5mutfdstk2	R5 IA8 SPOUSE OWNS FUNDS STOCK	7.1
18	ia5mutfdstk3	R5 IA8 SP SPOUSE OWN FUNDS STOCK	7.1
19	ia5mutfdstk4	R5 IA8 NO FUNDS OR STOCK OWNED	7.1
20	ia5ownbond1	R5 IA9 SP OWNS BONDS	7.6
21	ia5ownbond2	R5 IA9 SPOUSE OWNS BONDS	7.6
22	ia5ownbond3	R5 IA9 SP SPOUSE OWN BONDS	7.6
23	ia5ownbond4	R5 IA9 NO BONDS OWNED	7.6
24	ia5bnkaccdd1	R5 IA10 SP OWNS CHECK ACCT	5.6
25	ia5bnkaccdd2	R5 IA10 SPOUSE OWNS CHECK ACCT	5.6
26	ia5bnkaccdd3	R5 IA10 SP SPOUSE OWN CHECK ACCT	5.6
27	ia5bnkaccdd4	R5 IA10 NO CHECK ACCT OWNED	5.6
28	ia5bnkaccdd5	R5 IA10 SP OWNS SAVINGS ACCT	6.9
29	ia5bnkaccdd6	R5 IA10 SPOUSE OWNS SAVING ACCT	6.9
30	ia5bnkaccdd7	R5 IA10 SP SPOUSE OWN SAVNG ACT	6.9
31	ia5bnkaccdd8	R5 IA10 NO SAVINGS ACCT OWNED	6.9
32	ia5bnkaccdd9	R5 IA10 SP OWNS CDS	7.9
33	ia5bnkaccdd10	R5 IA10 SPOUSE OWNS CDS	7.9
34	ia5bnkaccdd11	R5 IA10 SP SPOUSE OWN CDS	7.9
35	ia5bnkaccdd12	R5 IA10 NO CDS OWNED	7.9
36	ia5realestt1	R5 IA13 SP OWNS REAL ESTATE	5.3
37	ia5realestt2	R5 IA13 SPOUSE OWNS REAL ESTATE	5.3
38	ia5realestt3	R5 IA13 SP SPOUSE OWN REAL ESTTE	5.3
39	ia5realestt4	R5 IA13 NO REAL ESTATE OWNED	5.3
40	lf5workfpay	R5 LF1 WORKED FOR PAY RECENTLY	2.9
41	lf5abstlstwk	R5 LF2 ABSENT FRM JOB LAST WEEK	7.4
42	lf5wrkplstmn	R5 LF3 WORK FOR PAY IN LST MONTH	0.0
43	lf5huswifwrk	R5 LF13 HUSB WIFE PARTN PAY WORK	3.2

Table 2. Round 5 Source Amount Variables

	Variable name	Label	% missing
1	ia5srrpymnt	R5 IA14 RECENT MTHLY SS RR PYMNT	14.2
2	ia5srrjtamt, ia5srrspamt, ia5srrptamt	IA14A-IA16A AMOUNT OF SOCIAL SECURITY/RAILROAD RETIREMNT	18.7
3	ia5srrjtest, ia5srrspest, ia5srrptest	IA14B-IA16B RANGE OF SOCIAL SECURITY/RAILROAD RETIREMNT	10.6
4	ia5ssipymnt	R5 IA17 RECENT MTHLY SSI PAYMENT	14.6
5	ia5ssijtamt, ia5ssispamt, ia5ssiptamt	IA17A-IA19A AMOUNT OF SUPPLEMENTAL SECURITY INCOME	22.9
6	ia5ssijtest, ia5ssispest, ia5ssiptest	IA17B-IA19B RANGE OF SUPPLEMENTAL SECURITY INCOME	13.0
7	ia5vapymnt	R5 IA20 RECENT MNTHLY VA PAYMENT	66.7
8	ia5vajtamt, ia5vaspamt, ia5vaptamt	IA20A-IA22A AMOUNT OF VETERANS ADMINISTRATION	22.0
9	ia5vajtest, ia5vaspest, ia5vapttest	IA20B-IA22B RANGE OF VETERANS ADMINISTRATION	14.0
10	ia5penpymnt	R5 IA23 RCNT MTH JOBPENSION PYMT	19.8
11	ia5penjtamt, ia5penspamt, ia5penptamt	IA23A-IA25A AMOUNT OF PENSION PLAN	22.8
12	ia5penjtest, ia5penspest, ia5penptest	IA23B-IA25B RANGE OF PENSION PLAN	13.2
13	ia5retworth	R5 IA26 RETIREMENT ACCOUNT WORTH	32.9
14	ia5retjtwrt, ia5retspwrt, ia5retptwrt	IA26A-IA28A AMOUNT OF ANY RETIREMENT ACCOUNTS	45.7
15	ia5retjtest, ia5retspest, ia5retptest	IA26B-IA28B RANGE OF ANY RETIREMENT ACCOUNTS	24.2
16	ia5rtlmwdrw	R5 IA29 RETRMNT WDRW AMT LST MTH	19.4
17	ia5rtlmjtwdr, ia5rtlmspwdr, ia5rtlmptwdr	IA29A-IA31A AMOUNT OF RETIREMENT ACCTS WITHDRAW LST MTH	21.2
18	ia5rtlmjtest, ia5rtlmspest, ia5rtlmptest	IA29B-IA31B RANGE OF RETIREMENT ACCTS WITHDRAW LST MTH	15.0
19	ia5rtyrwdrw	R5 IA32 RETRMNT WDRWL AMT LST YR	23.4
20	ia5rtyrjtamt, ia5rtyrspamt, ia5rtyrptamt	IA32A-IA34A AMOUNT OF PARTNERS ACCTS WITHDRAW LST MTH	30.6
21	ia5rtyrjtest, ia5rtyrspest, ia5rtyrptest	IA32B-IA34B RANGE OF PARTNERS ACCTS WITHDRAW LST MTH	19.0
22	ia5skbdwrth	R5 IA35 NONRETR STKS BNDS WRTH	36.5
23	ia5skbdjtwrt, ia5skbdspwrt, ia5skbdptwrt	IA35A-IA37A AMOUNT OF WORTH OF FUNDS & STOCKS	47.9
24	ia5skbdjtest, ia5bndjtest, ia5skbdspest, ia5bndspest, ia5skbdptest, ia5bndptest	IA35B-IA37C RANGE OF WORTH OF FUNDS & STOCKS	27.6
25	ia5bkcdwrth	R5 IA38 BANK ACCT CD WORTH	26.5
26	ia5bkcdjtwrt, ia5bkcdspwrt, ia5bkcdptwrt	IA38A-IA40A AMOUNT OF WORTH OF BANK ACCDS & CDS	34.3
27	ia5bkcdjtest, ia5bnkjtest, ia5bkcdspest, ia5bnkspest, ia5bkcdptest, ia5bnkptest	IA38B-IA40C RANGE OF WORTH OF BANK ACCDS & CDS	20.0
28	ia5itdvinc	R5 IA41 AMT INT DIV INCOME LS YR	36.5
29	ia5itdvjtamt, ia5itdvspamt, ia5itdvptamt	IA41A-IA43A AMOUNT OF TOT WRTH COMBNED INTERST	44.4
30	ia5itdvjtest, ia5itdvspest, ia5itdvptest	IA41B-IA43B RANGE OF TOT WRTH COMBNED INTERST	25.1
31	ia5brewrt	R5 IA44 BUSINESS REALESTATE WRTH	21.7
32	ia5brejtwrt, ia5brespwrt, ia5breptwrt	IA44A-IA46A AMOUNT OF TOT WRTH ALL REAL ESTATE	33.5
33	ia5brejtest, ia5brespest, ia5breptest	IA44B-IA46B RANGE OF TOT WRTH ALL REAL ESTATE	18.3
34	ia5breiinc	R5 IA47 BUS REALESTATE INC LS YR	22.1
35	ia5breijtamt, ia5breispamt, ia5breiptamt	IA47A-IA49A AMOUNT OF COMB INCOME FROM REL ESTAT	27.2
36	ia5breijtest, ia5breispest, ia5breiptest	IA47B-IA49B RANGE OF COMB INCOME FROM REL ESTAT	18.0
37	lf5ernfrmwrk	R5 LF10 AMT EARN FRM WRK LST MTH	7.0
38	lf5huwpaearn	R5 LF14 HUS/WIFE/PAR PAY LST MTH	39.9

Table 3. Auxilliary Variables

	Variable name	Label	% Missing
1	spmaritalstatus	R5 D MARITAL STATUS AT R5 INTERVIEW	0.1
2	pa5workfrpay	R5 PA17 EVER WORK FOR PAY	0.1
3	el5dhigstsch ¹	R5 EL10 D HGHST DGREE SCOOOL COMPLD	4.0
4	rl5dracehisp	R5 D RACE AND HISPANIC ETHNICITY	-
5	va5serarmfor	R5 VA1 SERVED IN ARMED FORCES	2.5
6	va5memnatgrd	R5 VA3 MEMBER OF NATIONAL GUARD	2.6
7	lf5mrthnonjb	R5 LF4 MOR THN ONE JOB LAST WEEK	-
8	lf5hrswkwork	R5 LF5 HRS PR WEEK WORK MAIN JOB	0.8
9	lf5hrwrkltwk	R5 LF6 HOURS WORK LAST WEEK	0.6
10	lf5hrwrklstw	R5 LF7 HOW MNY HOURS DID YOU WRK	2.2
11	lf5oftpaid	R5 LF8 HOW OFTN PAID ON MAIN JOB	0.3
12	lf5lstpaychk	R5 LF9 HOW MUCH LAST PAYCHECK	16.9
13	hp5ownrenttot	R5 HP1 OWN RENT OR OTHER	3.3
14	ia5howrecessp	R5 IA2 HOW REC SOC SEC PAY	0.9
15	ia5mstrecss	R5 IA3A MTH STRD REC SOC SEC PAY	31.5
16	ia5yrstrecsp	R5 IA3B YR STRTD REC SOC SEC PAY	14.5
17	ir5areacond1	R5 IR15 LITTER GLASS ON SDWLK ST	-
18	ir5areacond2	R5 IR15 GRAFFITI ON BLDG WALLS	-
19	ir5areacond3	R5 IR15 VACANT HOUSES OR STORES	-
20	ir5condhome1	R5 IR16 BROKEN WINDOWS IN HOME	-
21	ir5condhome2	R5 IR16 CRUMBLNG FOUNDTN IN HOME	-
22	ir5condhome3	R5 IR16 MISSNG BRCKS SIDNG IN HM	-
23	ir5condhome4	R5 IR16 ROOF PROBLEM IN HOME	-
24	ir5condhome5	R5 IR16 BROKEN STEPS TO HOME	-
25	ir5condhome6	R5 IR16 CONTINUOUS SIDEWALKS	-
26	fl5facility	R5 F ROUTING FLAG FROM RE4f HT3 5 6 7	-
27	Sex	HISKEW GENDER	-
28	agecat_r5	HISKEW AGE CATEGORY as of 2014	-
29	per_cap_inc_5yr	PER CAPITA INCOME [R5WT1; ACS]	-
30	lf5occupaton	R5 LF11 OCCUPATION MOST OF LIFE	4.9
31	lf5diffwrknw	R5 LF12 DO THS WRK SOMETHNG DIFF	0.2
32	smptype	O=Original smp, R=Replenishment smp	-

¹Equal to EL1HIGSTSCH for original sample cases and EL5HIGSTSCH for replenishment sample cases.

NOTE: - indicates the variable was never missing. The following auxiliary variables were included in Round 1 imputation but their corresponding variables were not included in Round 5 imputation: rl1yourrace1-8 and rtirace (not included because the derived variable rl5dracehisp was used instead); hh1livwthspo and hh1placekind (were skip pattern controllers for LF13 in Round 1, but not in Round 5).