NATIONAL HEALTH AND AGING TRENDS STUDY (NHATS)
Round 9 Income Imputation

September 9, 2020

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, 5, 7 and 9 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
Rounds 1 and 5 of NHATS collected information on sources of income (yes/no) and amounts for each source. Rounds 3, 7 and 9 also collected information on sources of income (yes/no) but not amounts for each source. Respondents with a spouse/partner were given the option of reporting sources for themselves either together or separately from their spouse/partner. Table 1 shows the income sources included in NHATS Round 9.

Table 1. Summary of Income Sources Collected in NHATS Round 9

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

--Asked about the existence of relevant asset

Extent of Missing Data for Total Income
Sixty-nine percent of the sample provided a total income amount and an additional 18% reported a bracketed value instead. Thus, a bracketed value could be created from reports for 87% of the sample and required imputation for 13%. An exact value was then imputed for 31% of the sample (18% within a reported bracketed value and 13% 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. Auxiliary variables that included respondent characteristics (e.g., age, race/ethnicity, gender, educational attainment, home ownership (in Round 9), veteran's status, labor force status (in Round 9), spouse/partner's labor force status (in Round 9), interviewer observations about the home condition and skip pattern controllers; and
3. Total income (reported or imputed), as well as source variables, from Round 7.

In order to preserve the joint distribution of the full set of income variables, all missing fencepost, auxiliary, and total income variables were imputed.

When imputing total income variables, both the Round 7 and Round 9 source variables and Round 7 total income 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
2018 Total Income Value		
Ia9totinc	R9 IA50 TOTAL INCOME	Actual reported $ amounts
Ia9toincimf	R9 F IMPUTED TOTAL INC FLG	Flag indicating imputation
		Imputed values 1-5 for missing $
		amounts and reported bracket
Ia9toincim1-5	R9 IA50 IMPUTED TOTAL INC1-INC5	amounts
2018 Total Income Range – Respondents who have spouse/partner		
Ia9toincesjt	R9 IA51A JOINT EST TOT INCOME	Actual reported $ amounts
Ia9eincimjf	R9 F IMPTD JOINT EST TOT INC FLG	Flag indicating imputation
		Imputed values 1-5 for missing $
		amounts and reported bracket
Ia9eincimj1-5	R9 IA51A IMP EST JOINT TOT INC1-INC5	amounts
2018 Total Income Range – Respondents who are single		
Ia9toincessg	R9 IA51B SNGLE EST TOT INC	Actual reported $ amounts
Ia9eincimsf	R9 F IMPUTED SGL EST TOT INC FLG	Flag indicating imputation
		Imputed values 1-5 for missing $
		amounts and reported bracket
Ia9eincims1-5	R9 IA51B IMP EST SP SGL TOT INC1-INC5	amounts

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 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. In Round 9, item respondents with actual reported $ amounts were not included in the 5 sets of imputed variables.

Because Round 7 variables were used in the imputation of Round 9 variables, in order to capture the effects of imputation of the Round 7 variables on the precision of estimates involving the Round 9 variables, the five sets of imputed values for the Round 7 variables were used to impute the five sets of imputed values for the Round 9 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 9 Income Imputation

Table 1. Round 9 Source ("Fencepost") Variables

#	Variable name	Label	% missing
1	ia9recsspa1	R9 IA1 SP REC SOCIAL SECURITY	1.3
2	ia9recsspa2	R9 IA1 SPOUSE PART REC SOC SEC	1.3
3	ia9recsspa3	R9 IA1 NO SOC SECURTY PAYMNT REC	1.3
4	ia9recssils1	R9 IA4 SP RECEIVD SSI LAST MONTH	2.4
5	ia9recssils2	R9 IA4 SPOUSE PRT RECSSI LST MO	2.4
6	ia9recssils3	R9 IA4 NO SSI RECEIVD LAST MONTH	2.4
7	ia9rvapayls1	R9 IA5 SP REC PAY FRM VA LAST MO	1.1
8	ia9rvapayls2	R9 IA5 SPOUS PA REC VA PAY LSTMO	1.1
9	ia9rvapayls3	R9 IA5 NO VA PAY REC LAST MONTH	1.1
10	ia9penjobou1	R9 IA6 SP HAS PENSION PLAN	1.6
11	ia9penjobou2	R9 IA6 SPOUSE HAS PENSION PLAN	1.6
12	ia9penjobou3	R9 IA6 NO PENSION PLAN	1.6
13	ia9iraothac1	R9 IA7 SP HAS IRA OTH RETIRE ACC	3.7
14	ia9iraothac2	R9 IA7 SPOUSE HAS IRA OTHR ACC	3.7
15	ia9iraothac3	R9 IA7 NO IRA OTHR RETIRE ACCT	3.7
16	ia9mutfdstk1	R9 IA8 SP OWNS MUTUAL FUND STOCK	3.0
17	ia9mutfdstk2	R9 IA8 SPOUSE OWNS FUNDS STOCK	3.0
18	ia9mutfdstk3	R9 IA8 SP SPOUSE OWN FUNDS STOCK	3.0
19	ia9mutfdstk4	R9 IA8 NO FUNDS OR STOCK OWNED	3.0
20	ia9ownbond1	R9 IA9 SP OWNS BONDS	3.6
21	ia9ownbond2	R9 IA9 SPOUSE OWNS BONDS	3.6
22	ia9ownbond3	R9 IA9 SP SPOUSE OWN BONDS	3.6
23	ia9ownbond4	R9 IA9 NO BONDS OWNED	3.6
24	ia9bnkacccd1	R9 IA10 SP OWNS CHECK ACCT	1.8
25	ia9bnkacccd2	R9 IA10 SPOUSE OWNS CHECK ACCT	1.8
26	ia9bnkacccd3	R9 IA10 SP SPOUSE OWN CHECK ACCT	1.8
27	ia9bnkacccd4	R9 IA10 NO CHECK ACCT OWNED	1.8
28	ia9bnkacccd5	R9 IA10 SP OWNS SAVINGS ACCT	2.6
29	ia9bnkacccd6	R9 IA10 SPOUSE OWNS SAVING ACCT	2.6
30	ia9bnkacccd7	R9 IA10 SP SPOUSE OWN SAVNG ACT	2.6
31	ia9bnkacccd8	R9 IA10 NO SAVINGS ACCT OWNED	2.6
32	ia9bnkacccd9	R9 IA10 SP OWNS CDS	4.1
33	ia9bnkacccd10	R9 IA10 SPOUSE OWNS CDS	4.1
34	ia9bnkacccd11	R9 IA10 SP SPOUSE OWN CDS	4.1
35	ia9bnkacccd12	R9 IA10 NO CDS OWNED	4.1
36	ia9realestt1	R9 IA13 SP OWNS REAL ESTATE	1.4
37	ia9realestt2	R9 IA13 SPOUSE OWNS REAL ESTATE	1.4
38	ia9realestt3	R9 IA13 SP SPOUSE OWN REAL ESTTE	1.4
39	ia9realestt4	R9 IA13 NO REAL ESTATE OWNED	1.4
40	lf9workfpay	R9 LF1 WORKED FOR PAY RECENTLY	0.1
41	lf9abstlstwk	R9 LF2 ABSENT FRM JOB LAST WEEK	0.1
42	lf9wrkplstmn	R9 LF3 WORK FOR PAY IN LST MONTH	-[1]
43	lf9huswifwrk	R9 LF13 HUSB WIFE PARTN PAY WORK	0.4

---

[1] - indicates the variable was never missing.

Table 2. Auxiliary Variables

	Variable name	Label	% Missing
1	sex	GENDER	-
2	agecat_r5	AGE CATEGORY AS OF 2014	-
3	rtirace	RACEETH, 3-CATEGORY	-
4	per_cap_inc_5yr	PER CAPITA INCOME [ACS]	-
5	smptype	SAMPLE TYPE (O=ORIGINAL SMP, R=REPLENISHMENT SMP)	-
6	el5dhigstsch[2]	R5 EL10 D HGHST DGREE SCOOL COMPLD	-
7	rl5dracehisp	R5 D RACE AND HISPANIC ETHNICITY	-
8	va5serarmfor	R5 VA1 SERVED IN ARMED FORCES	-
9	va5memnatgrd	R5 VA3 MEMBER OF NATIONAL GUARD	-
10	fl9facility	R9 F ROUTING FLAG FROM RE4f HT3 5 6 7	-
11	ir9areacond1	R9 IR15 LITTER GLASS ON SDWLK ST	-
12	ir9areacond2	R9 IR15 GRAFFITI ON BUILDG WALLS	-
13	ir9areacond3	R9 IR15 VACANT HOUSES OR STORES	-
14	ir9condhome1	R9 IR16 BROKEN WINDOWS IN HOME	-
15	ir9condhome2	R9 IR16 CRUMBLNG FOUNDTN IN HOME	-
16	ir9condhome3	R9 IR16 MISSNG BRCKS SIDNG IN HM	-
17	ir9condhome4	R9 IR16 ROOF PROBLEM IN HOME	-
18	ir9condhome5	R9 IR16 BROKEN STEPS TO HOME	-
19	ir9condhome6	R9 IR16 CONTINUOUS SIDEWALKS	-
20	hh9dmarstat	R9 D MARITAL STATUS AT R9	-
21	hh9livwthspo	R9 HH11 LIVE WITH SPOUSE PARTNER	0.1
22	hh9placekind	R9 HH12 KIND OF PLACE LIVE IN	1.8
23	pa9workfrpay	R9 PA17 EVER WORK FOR PAY	0.1
24	lf9mrthnonjb	R9 LF4 MOR THN ONE JOB LAST WEEK	0.2
25	lf9hrswkwork	R9 LF5 HRS PR WEEK WORK MAIN JOB	1.0
26	lf9hrwrkltwk	R9 LF6 HOURS WORK LAST WEEK	0.9
27	lf9hrwrklstw	R9 LF7 HOW MNY HOURS DID YOU WRK	7.7
28	lf9oftpaid	R9 LF8 HOW OFTN PAID ON MAIN JOB	-
29	hp9ownrentot	R9 HP1 OWN RENT OR OTHER	0.6

---

[2] Equal to EL1HIGSTSCH for original sample cases and EL5HIGSTSCH for replenishment sample cases.

Table 3. Round 7 Source ("Fencepost") Variables[3]

#	Variable name	Label
1	ia7recsspa1	R7 IA1 SP REC SOCIAL SECURITY
2	ia7recsspa2	R7 IA1 SPOUSE PART REC SOC SEC
3	ia7recsspa3	R7 IA1 NO SOC SECURTY PAYMNT REC
4	ia7recssils1	R7 IA4 SP RECEIVD SSI LAST MONTH
5	ia7recssils2	R7 IA4 SPOUSE PRT REC SSI LST MO
6	ia7recssils3	R7 IA4 NO SSI RECEIVD LAST MONTH
7	ia7rvapayls1	R7 IA5 SP REC PAY FRM VA LAST MO
8	ia7rvapayls2	R7 IA5 SPOUS PA REC VA PAY LSTMO
9	ia7rvapayls3	R7 IA5 NO VA PAY REC LAST MONTH
10	ia7penjobou1	R7 IA6 SP HAS PENSION PLAN
11	ia7penjobou2	R7 IA6 SPOUSE HAS PENSION PLAN
12	ia7penjobou3	R7 IA6 NO PENSION PLAN
13	ia7iraothac1	R7 IA7 SP HAS IRA OTH RETIRE ACC
14	ia7iraothac2	R7 IA7 SPOUSE HAS IRA OTHR ACC
15	ia7iraothac3	R7 IA7 NO IRA OTHR RETIRE ACCT
16	ia7mutfdstk1	R7 IA8 SP OWNS MUTUAL FUND STOCK
17	ia7mutfdstk2	R7 IA8 SPOUSE OWNS FUNDS STOCK
18	ia7mutfdstk3	R7 IA8 SP SPOUSE OWN FUNDS STOCK
19	ia7mutfdstk4	R7 IA8 NO FUNDS OR STOCK OWNED
20	ia7ownbond1	R7 IA9 SP OWNS BONDS
21	ia7ownbond2	R7 IA9 SPOUSE OWNS BONDS
22	ia7ownbond3	R7 IA9 SP SPOUSE OWN BONDS
23	ia7ownbond4	R7 IA9 NO BONDS OWNED
24	ia7bnkacccd1	R7 IA10 SP OWNS CHECK ACCT
25	ia7bnkacccd2	R7 IA10 SPOUSE OWNS CHECK ACCT
26	ia7bnkacccd3	R7 IA10 SP SPOUSE OWN CHECK ACCT
27	ia7bnkacccd4	R7 IA10 NO CHECK ACCT OWNED
28	ia7bnkacccd5	R7 IA10 SP OWNS SAVINGS ACCT
29	ia7bnkacccd6	R7 IA10 SPOUSE OWNS SAVING ACCT
30	ia7bnkacccd7	R7 IA10 SP SPOUSE OWN SAVNG ACT
31	ia7bnkacccd8	R7 IA10 NO SAVINGS ACCT OWNED
32	ia7bnkacccd9	R7 IA10 SP OWNS CDS
33	ia7bnkaccc10	R7 IA10 SPOUSE OWNS CDS
34	ia7bnkaccc11	R7 IA10 SP SPOUSE OWN CDS
35	ia7bnkaccc12	R7 IA10 NO CDS OWNED
36	ia7realestt1	R7 IA13 SP OWNS REAL ESTATE
37	ia7realestt2	R7 IA13 SPOUSE OWNS REAL ESTATE
38	ia7realestt3	R7 IA13 SP SPOUSE OWN REAL ESTTE
39	ia7realestt4	R7 IA13 NO REAL ESTATE OWNED
40	lf7workfpay	R7 LF1 WORKED FOR PAY RECENTLY
41	lf7abstlstwk	R7 LF2 ABSENT FRM JOB LAST WEEK
42	lf7wrkplstmn	R7 LF3 WORK FOR PAY IN LST MONTH
43	lf7huswifwrk	R7 LF13 HUSB/WIFE/PARTN PAY WORK
44	ia7totinc	R7 IA50 TOTAL INCOME

---

[3] Round 7 source variables are not available for 16 R7 respondents who had completed facility questionnaires only, so their imputation process only used Table 1 and Table 2 variables.