Codebook for SIPP Synthetic Beta File

This codebook documents version 5.1 of the SIPP Synthetic Beta (SSB). The SSB is a set of files containing individual-level data synthesized from linked survey and administrative data. The SSB is produced by the US Census Bureau as part of a joint project with the Social Security Administration (SSA), and the Internal Revenue Service (IRS). The goal of the project is to make some of the benefits of linked survey and administrative data available to researchers outside of restricted-access Census Bureau facilities in a manner that protects the confidentiality of the underlying data.

Creation of the SSB

The SSB is created from several data sources. The survey data are drawn from multiple panels of the Survey of Income and Program Participation (SIPP): the 1990, 1991, 1992, 1993, 1996, and 2004 panels. The administrative data are drawn from the following SSA files: the Master Earnings File, the Master Beneficiary Records (MBR), the Supplemental Security Records (SSR), the 831 Disability File (F831), and the Payment History Update System (PHUS).

The creation of the SSB begins with the construction of the Gold Standard File (GSF). To construct the GSF, a set of variables from the 1990-2004 SIPP panels are standardized to produce consistent measures across panels. The SIPP respondent identifiers are mapped to Social Security Numbers (SSB) using the Census Bureau's erson Information Validation System (PVS). Using the list of SSN's for the sample is SIPP respondents, SSA creates Summary Earnings Records (SER) and Detailed Earnings Record (DER) extracts from the Master Earnings File. SSA also creates extracts from the four benefit files (MBR, SSR, F831, and PHUS) from the corresponding master files. Using the mapping between the SIPP identifiers and SSN's, Census then links these extracts to the SIPP data. The GSF consists of person-level research variables created from these linked data.

The next step in the creation of the SSB is to impute missing values in the GSF multiple times. This process results in four files (implicates) referred to as the Completed Data implicates. Each of these implicates contains original GSF values where non-missing and imputed values where the original value is missing. The imputations across Completed Data implicates are independent of each other.

The Completed Data implicates from the basis of the data synthesis that produces the SSB files. From each Completed Data file, four synthetic datasets are created by synthesizing variables conditional on the values in the Completed Data file. Thus, the SSB consists of sixteen files (implicates). All but the following data are synthesized in the SSB implicates: gender, OASDI benefit type, and spouse link (specific variables described in the data items section below). Detailed documentation of the process of data synthesis is available in the publication "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project" which can be downloaded from www.census.gov/sipp/syth_data.html.

The Completed Data and SSB implicates need not all have the same number of records. In order to be included in a Completed Data or SSB implicate, an individual's (possibly imputed or synthesized) age must be at least fifteen years as of January 1 in the first year of his or her SIPP panel. The interaction between this restriction and the variation in imputed and synthesized ages across implicates causes the exclusion of a slightly

different set of individuals from each Completed Data and SSB implicate.

Using SSB

The GSF and Completed Data implicates contain personally identifiable information protected by Titles 13, 26, and 42 and cannot be accessed without Census Bureau Special Sworn Status nor outside of Census Bureau facilities. The SSB files, however, havve been cleared by the Census Bureau Disclosure Review Board, SSA, and IRS for use by individuals without Census Bureau Special Sworn Status and outside of Census Bureau facilities.

Researchers interested in using the SSB can submit an application to the Census Bureau. The application form and instructions can be downloaded from www.census.gov/sipp/syth_data.html. Applications will be judged solely on feasability of the proposed project (i.e., that the necessary variables are available on the SSB). Once an application has been accepted, the new user will be given an account on a server where the data can be accessed and analyzed. While no SSB data downloads are permitted at this time, users do not have to operate behind the Census Bureau firewall to access this server.

The SSB is designed to be analytically valid in that sense that point estimates should be unbiased and estimated variances should lead to inferences similar to those that would be drawn from an identical analysis on the Completed Data implicates. Initial tests of analytic validity of the SSB have been promising. All SSB users are invited to help further test the analytic validity of the SSB by submitting programs used to analyze the SSB to be run on the Completed Data and/or Gold Standard files. Users need only inform Census Bureau staff of the location on the server of such programs and work with Census Bureau staff to ensure that the programs run without error. Census Bureau staff will run the programs on the confidential data and release to the user resulting output that are cleared for release by the Census Bureau Disclosure Review Board. In order to evaluate the effects of the data synthesis separate from the effect of imputing missing data, comparisons should be made between results from the SSB and the Completed Data. To evaluate the effects of missing data imputation, comparisons should be made between results from the Completed Data and the Gold Standard.

When analyzing the SSB, users should account for the multiple imputation aspect of the SSB by averaging statistics of interests across all sixteen implicates. Variance measures should be created following the appropriate multiple imputation formulae as described in the document "Using the SIPP Synthetic Beta for Analysis" which can be downloaded www.census.gov/sipp/synth_data.html.

Description of Data Items

The remainder of this document lists and describes the data items contained in the SSB files. For each data item, the variable name, variable type and length, source, range of values, and description are provided. For data items for which the universe depends on another data item, the parent variable and the values that the parent variable must take in order for the data item to be in-scope are also listed.

Any questions related to the SSB can be directed to sehsd.synthetic.data.use.list@census.gov.

---------- Identifiers ----------

---

panel SIPP Panel Year

type: numeric (double)

range: [1990,2004]

Indicates panel of source record

---

personid Unique person identifier

type: numeric (double)

range: positive values

Across the Gold Standard and Completed Data files, personid uniquely identifies SIPP respondents. In the SSB, personid uniquely identifies records within a particular implicate. In order to strengthen confidentiality protection, personid in the SSB does not link records across implicates or to the Gold Standard and Completed Data files.

---

spouse_personid Personid of spouse

type: numeric (double)

range: positive values

Personid of linked spouse. Across the Gold Standard and Completed Data files, spouse_personid uniquely identifies spouses of SIPP respondents. In the SSB, spouse_personid uniquely identifies records within a particular implicate. In order to strengthen confidentiality protection, spouse_personid in the SSB does not link records across implicates or to the Gold Standard and Completed Data files.
Linked spouse is defined as the first person to whom the SIPP respondent was married during the time period covered by the SIPP panel. Individuals could enter the panel already married and then each would be linked to the other. Individuals could also get married during the course of the panel. If this was the first observed marriage for each member of the couple, they were linked together. Individuals could also get divorced during the course of the panel and then remarry. In many cases, this later marriage caused a new individual to join the panel. This new SIPP respondent would only be linked to his or her spouse if the spouse (and original SIPP sample member) had not already been observed married to someone else. If the original SIPP sample member had been previously linked by marriage to another SIPP sample member, this original link was maintained in spouse_personid. However the marital history reflects the ending of this marriage and the occurrence of the next marriage for the original SIPP sample member. Likewise, the new SIPP sample member who joins through marriage will have that marriage date recorded in his or her marital history but will have a blank spouse_personid.
In summary, this variable captures only one marriage partner and does not provide a history of marriage partners even if this history is (partially) observed in the SIPP.
The link between SIPP respondents and their spouses has not been perturbed in any way in the SSB. The same individuals will be linked as married partners in the Gold Standard, the Completed Data, and the SSB.

Demographic Variables

The variables in this section are all drawn from the SIPP and represent demographic information gathered by the survey at a specific point in time. Entries for individual variables describe the exact SIPP source variable and the reference point in time.

---

male — Male

type: numeric (double)

range: [0,1]

Numeric	Label
0	Female
1	Male

This variable tells whether a person is male or female

---

race — Race

type: numeric (double)

range: [1,3]

Numeric	Label
1	White
2	Black
3	Other

In the 1984 and 1990-2008 Census-internal SIPP panels, a value for race is included on each wave file. Thus, there are actually as many race variables as there are waves of the survey and some changes occur across waves as a result of data collection error. Race is chosen by creating an array of variables race1-race{max number of waves} and choosing the first non-missing value. Thus race comes from the first wave in which the individual was interviewed instead of from a fixed point in the survey.

---

hispanic — Hispanic

type: numeric (double)

range: [0,1]

Numeric	Label
0	Non-hispanic
1	Hispanic

In the 1984 and 1990-1993 SIPP panels, a value for ethnicity is included on each wave file. Thus, there are actually as many ethnicity variables

as there are waves of the survey and some changes occur across waves as a result of data collection error. Ethnicity is chosen by creating an array of variables ethncty1-ethncty{max number of waves} and choosing the first non-missing value. Thus, ethnicity comes from the first wave in which the individual was interviewed instead of from a fixed point in the survey. Respondents are coded as Hispanic if they have an ethnicity code between 14 and 20. In the 1996-2008 panels, the longitudinally-edited version contains only one value for ethnicity across all waves (eorigin) and this value is used. Respondents are coded as Hispanic if they have an ethnicity code between 20 and 28 in 1996 and 2001, or if they have an ethnicity code of 1 in 2008.

---

foreign_born — Foreign Born

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Born in U.S.
1	Born in country other than U.S.

Immigrant Status, born in counrty other than U.S.
Taken from wave 2 topical module (TM8730, TM8734, TM8709 1984 and 1990-1993 panels; eprstate, ebrstate and rcitiznt 1996 panel; eprstate, ebrstate and tcitiznt 2001 panel; eprstate, ebrstate, citiz, and ebornus panel)

---

time_arrive_usa — Decade of Arrival to US (Foreign Born)

---

type: numeric (double)

range: [1,10]

Decade arrive in U.S. (answered when SIPP respondent was foreign_born)
The year of arrival to the U.S. is from the Census-internal SIPP files (TM8736 1984 and 1990-1993 panels, rmoveus 1996 panel; tmoveus 2001-2008 panels)

---------- Disability Variables ----------

---

sum_disab — Disability (Sum of Core and TM)

---

type: numeric (double)

range: [0,1]

Health limits kind or amount of work
For the 1996-2004 panels, information on work-limiting disability comes

from core question (edisabl) during wave 2 for people ages 15-69, when respondents were asked whether health limited the type or amount of work. This indicator is supplemented with details from the Functional Limitations and Disability topical module (wave 5 for 1996-2004 panels, variable ejobdif) that covers people ages 16-67. For the 1990-1993 panels, disability information comes from the core question (disab) during wave 2 for people ages 15-69, when respondents were asked whether health limited the type or amount of work. This information is supplemented with details from the Functional Limitations and Disability topical module (waves 3, 3, 6, 3 for 1990-1993 panels, variables TM8914, TM8918, and TM8920) that covers people ages 16-67. In order to make the sum_disab consistent for all panels, both responses were taken from the core and topical modules, when available, with any positive indication of health limiting the kind or amount of work flagging the positive response.

---

sum_disab_nowork — Disability Prevents Work (Sum of Core and TM)

---

type: numeric (double)

range: [0,1]

In scope to be asked questions about whether health prevents work.

Education Variables

---

year_beg_posths — Year Began Post-HS Education

---

type: numeric (double)

range: YYYY

The wave 2 Education and Training History topical module povides knowledge of the year that post-high school education began (variable TM8420 for 1984 and 1990-1993 panels; variable tcollstr for 1996-2008 panels).

---

current_enroll_coll — Flag currently enrolled in college

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Not currently enrolled in college
1	Currently enrolled in college

Indicates whether an individual is enrolled in college at time of SIPP education history topical module and has not finished his/her education.

This variable can be used to differentiate between individuals who completed some college and stopped school and those who have finished some college but not yet stopped attending school.

---

year_bach — Year of Bachelor Degree

---

type: numeric (double)

range: YYYY

Year bachelor's degree was finished

---

field_bach — Field of Bachelor Degree

---

type: numeric (double)

range: [0,20]

Numeric	Label
0	
1	Agriculture/Forestry
2	Art/Architecture (1996-2004 panels); Biology (1990-1993 panels)
3	Business/Management
4	Communications (1996-2004 panels); Economics (1990-1993 panels)
5	Computer and Information Sciences (1996-2004 panels); Education (1990-1993 panels)
6	Education (1996-2004 panels); Engineering (1990-1993 panels)
7	Engineering (1996-2004 panels); English/Journalism (1990-1993 panels)
8	English/Literature (1996-2004 panels); Home Economics (1990-1993 panels)
9	Foreign Language (1996-2004 panels); Law (1990-1993 panels)
10	Health Sciences (1996-2004 panels); Liberal Arts/Humanities (including arts, architecture, music, languages, philosophy) (1990-1993 panels)
11	Liberal Arts/Humanities (1996-2004 panels); Mathematics/Statistics (1990-1993 panels)
12	Mathematics/Statistics (1996-2004 panels); Medicine (1990-1993 panels)
13	Nature Sciences (Biological and Physical) (1996-2004 panels); Nursing, Pharmacy, Health

	Technologies (1990-1993)
14	Philosophy/Religion/Theology (1996-2004 panels); Physical or Earth Sciences (1990-1993 panels)
15	Pre-Professional (1996-2004 panels); Police Science or Law Enforcement (1990-1993 panels)
16	Psychology
17	Social Sciences/History (1996-2004 panels); Religion/Theology (1990-1993 panels)
18	Other (1996-2004 panels); Social Sciences (1990-1993 panels)
19	Vocational or Technical Studies (1990-1993 panels)
20	Other (1990-1993 panels)

Field in which bachelor's degree was obtained. Taken from topical history module on education history.

---

year_end_posths — Year Ended Post-HS Education

---

type: numeric (double)

range: YYYY

The wave 2 Education and Training History topical module provides knowledge of the year that post-high school education ended (variables TM8426 and TM8440 for 1984 and 1990-1993 panels; variables tlastcol, tvocyr, tassocyr, tbachyr, and tadvncyr for 1996-2008 panels

---

educ_5cat — Education Category (5)

---

type: numeric (double)

range: [1,5]

Numeric	Label
1	No high school degree
2	High school degree
3	Some college
4	College degree
5	Graduate degree

Highest level of education attained at the time of the education history topical module. This variable was created from information gathered in the topical module on education history and represents the highest level of education achieved up to the point of the adminstration of the topical module questions.
For individuals who did not answer the topical module eduation history questions, we created a highest level of education variable from the SIPP core and used this variable as a predictor to impute the education variable based on the topical module. See description of educ_cat for details on how the SIPP core education variable was created.

---

current_enroll_hs — Flag currently enrolled in high school

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Not currently enrolled in high school
1	Currently enrolled in high school

Indicates whether individual is still enrolled in high school (or lower grade) and has not completed education at time of SIPP education history topical module

---

year_ended_hs — Year Ended HS (or less) Education

---

type: numeric (double)

range: YYYY

The wave 2 Education and Training History topical module provides knowledge of the year that high school was last attended (variables TM8404 and TM8412 for 1984 and 1990-1993 panels; variables tlstschl and thsyr for 1996-2008 panels).

Economic Variables

The variables in this section are all drawn from the SIPP and represent economic information gathered by the survey at a specific point in time. Entries for individual variables describe the exact SIPP source variable and the reference point in time.

---

totnetworth — Total Net Worth

---

type: numeric (double)

range: positive/negative values

Total net worth

---

pension_in_scope_empl — In-Scope for Pension (Level II)

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Pension was not in scope
1	Pension was in scope

Individual must have been employed at time of pension topical module in order to answer the pension questions.

---

dc_pension — Defined Contribution Pension Plan

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	No defined contribution pension plan
1	Had defined contribution pension plan

---

db_pension — Defined Benefit Pension Plan

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	No defined benefit pension plan
1	Had defined benefit pension plan

---

own_home — Own a Home

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Do not own a home
1	Own a home

---

homeequity — Home Equity

---

type: numeric (double)

range: positive/negative values

Self-reported home equity value

---

nonhouswealth — Non-Housing Financial Wealth

---

type: numeric (double)

range: positive/negative values

Non-housing wealth = total wealth minus home equity

---

ind_exist — Flag: Industry Assigned

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	No, last worked 1984 or earlier, or no valid industry reported
1	Yes

Does person have valid industry from a job held during survey

---

ind_4cat — Industry Category (4)

---

type: numeric (double)

range: [1,4]

Numeric	Label
1	Manufacturing
2	Wholesale/retail trade
3	FIRE, services, public administration, military
4	Agriculture, mining, construction, transportation, communications, and public utilities

Industry is a characteristic of an individual's job and hence varies over time. There are industry values reported for (potentially) two jobs in each wave of the survey. Industry is chosen by summing earnings associated with the array of variables ws1ind1-ws1ind{max number of waves} and ws2ind1-ws2{max number of waves} in the 1984 and 1990-1993 panels, and ejbind1_1-ejbind1_{max number of waves} and ejbind2_1-ejbind2_{max number of waves} in the 1996-2008 panels and choosing the industry associated with the greatest total earnings. Thus industry is the industry from which greatest earnings are derived in the survey.

---

occ_exist — Flag: Occupation Assigned

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	No, last worked 1984 or earlier, or no valid industry reported

1 Yes

Does person have valid occupation from a job held during survey

---

occ_3cat — Occupation Category (3)

---

type: numeric (double)

range: [1,3]

Numeric	Label
1	Managerial and professional specialty occupations
2	Technical, sales, and administrative support occupations
3	Other

Occupation is a characteristic of an individual's job and hence varies over time. There are occupation values reported for (potenitally) two jobs in each wave of the survey. Occupation is chosen by summing earnings associated with the array of variables ws1occ1-ws1occ{max number of waves} and ws2occ1-ws2occ{max number of waves} in the 1984 and 1990-1993 panels, and tjbocc1_1-tjbocc1_{max number of waves} and tjbocc2_1-tjbocc2_{max number of waves} in the 1996-2008 panels and choosing the occupation associated with the greatest total earnings. Thus occupation is the occupation from which greatest earnings are derived in the survey.

Marital History Variables

Marital history is presented as two arrays of 8 elements describing up to 4 marriages. This history retains at most 4 dates of origin and the associated 4 dates of dissolution (whether due to divorce or death), if applicable, along with the corresponding type of event (marriage, divorce, or widowhood) for each SIPP respondent. The wave 2 Marital History topical module provides the majority of this information for up to 3 marriages. If an individual had more than 3 marriages, no dates for those marriages between the second and most recent are collected during the topical module interview.

For individuals who participate in the topical module, we supplement this information by searching for new marriages or the termination of an existing marriage utilizing our knowledge of monthly marital status covering the period of the panel beyond wave 2. We rely exclusively on the complete set of monthly marital status indicators for people who do not participate in the topical module.

The two marital history arrays, one comprised of dates and the other of event types, are edited to ensure internal consistency for linked spouses. Missing event dates or reasons are acquired from the spouse who has provided this information during a SIPP interview. This means that in the case of a deceased individual, the surviving spouse's report of widowhood is transferred to their former spouse. Likewise, for respondents who leave the household due to divorce (and are no longer interviewed by the SIPP), the spouse remaining in the household supplies the details

regarding marital dissolution for both.

It is possible for beginning or ending date information to not identically match for the linked spouses. In this case, we evaluate whether topical module details were supplied. If both spouses participated in the topical module, the data are considered as likely to be reliable from either individual. We examine whether the first spouse's beginning date occurs before the previous marriage end date for either spouse while the second spouse's beginning date occurs after the previous marriage end date of both spouses. If so, then the second spouse's beginning date replaces the first spouse's beginning date. Alternatively, if the reverse is true, then the first spouse's beginning date replaces the second spouse's beginning date. When no obvious conflicts with the date of the start of the current marriage and the date of termination of the previous marriages of either spouse exist, then a random number is used to determine which spouse's information to retain. A similar algorithm is implemented to resolve issues relating to non-matching ending dates of the linked spouses (in this case, each spouse's subsequent marital start date is taken into consideration).

When disagreements between the beginning or ending dates occur and only one of the two spouses participated in the topical module, the participating spouse's information is considered to be more reliable (provided that the adoption of the date presents no conflicts with previous or subsequent marital events for either spouse). In the absence of topical module participation for both spouses, a random number is used to determine which spouse's information to retain. Any persisting conflicts between dates of marital events are remedied by utilizing a random number to determine the most reasonable event date for the pair. As a final step, the cleaned file is carefully reviewed once more to ensure internal consistency.

---

mh1 — Flag: Marital History Event 1

---

type: numeric (double)

range: [1,1]

Numeric	Label
1	First marriage occurred

---

mh2 — Flag: Marital History Event 2

---

type: numeric (double)

range: [1,2]

Numeric	Label
1	First marriage ended in widowhood
2	First marriage ended in divorce/separation

---

mh3 — Flag: Marital History Event 3

---

type: numeric (double)

range: [1,1]

Numeric	Label
1	Second marriage occurred

---

mh4 — Flag: Marital History Event 4

---

type: numeric (double)

range: [1,2]

Numeric	Label
1	Second marriage ended in widowhood
2	Second marriage ended in divorce/separation

---

mh5 — Flag: Marital History Event 5

---

type: numeric (double)

range: [1,1]

Numeric	Label
1	Third marriage occurred

---

mh6 — Flag: Marital History Event 6

---

type: numeric (double)

range: [1,2]

Numeric	Label
1	Third marriage ended in widowhood
2	Third marriage ended in divorce/separation

---

mh7 — Flag: Marital History Event 7

---

type: numeric (double)

range: [1,1]

Numeric	Label
1	Fourth marriage occurred

---

mh8 — Flag: Marital History Event 8

---

type: numeric (double)

range: [1,2]

Numeric	Label
1	Fourth marriage ended in widowhood
2	Fourth marriage ended in divorce/separation

---

flag_mar4t — Flag: 4 or More Marriages

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	No additional marriage occurred with unknown date
1	An additional marriage occurred but with unknown date

Flag for existence of a marriage for which date is unknown because it was not collected in the SIPP. The marital history topical module asks about a person's first and second marriages and then his or her most recent marriage. If any other marriages occurred after the second but before the most recent, no information about this marriage is collected. However, individuals are categorized as having 1, 2, 3, or more than 3 marriages. We create flag_mar4t to identify individuals who reported more than 3 marriages.

---

obs_first_sipp_mar_num — Ordinal Number of First Obs Marriage

---

type: numeric (double)

range: [1,4]

---

sipp_panel_end_date — SIPP Panel End Date

---

type: numeric (double)

range: positive values

---

mh_date — Date of Marital History Event <n>

---

type: numeric (double)

range: positive/negative values

---

cur_endmar_flag Flag: Linked marriage ended

type: numeric (double)

range: [0,1]

cur_endmar_reas Flag: Reason linked marriage ended

type: numeric (double)

range: [0,1]

cur_startmar SAS Date linked marriage began

type: numeric (double)

range: positive/negative values

cur_endmar SAS Date linked marriage ended

type: numeric (double)

range: positive values

Fertility Variables

Number of children and dates of birth

own_kids_ever Number of Children Ever Born

type: numeric (double)

range: [0,6]

Number of children ever born. This is taken from the wave 2 Fertility history topical module (TM8752 and TM8754 for 1984 and 1990-1993 panels; tfrchl and tmomchl for 1996-2008 panels).

first_birth_year Year of Birth of First Child

type: numeric (double)

range: YYYY

This is taken from the wave 2 Fertility history topical module (TM8762 and TM8794 for 1984 and 1990-1993 panels; tfbrthyr for 1996-2008 panels).

---

last_birth_year — Year of Birth of Last Child

---

type: numeric (double)

range: YYYY

This is taken from the wave 2 Fertility history topical module (TM8768 and TM8782 for 1984 and 1990-1993 panels; tlbirtyr for 1996-2008 panels).

---

totfam_kids — Total Number of Children in Family

---

type: numeric (double)

range: [0,12]

Number of children under the age of 18 that live in a family in the interview month in which marital status is first observed (for those without spouses during the course of the SIPP) or in which the respondent's spouse is assigned. This number is the same for all family members and does not indicate that the children are related to a particular individual (fnkids for 1984 and 1990-1993 panels, rfnkids for 1996-2008 panels).

Lifespan Variables

Birth and Death dates

---

birthdate — Date of Birth

---

type: numeric (double)

range: positive/negative values

This variable was taken from a hierarchy of SSA sources instead of the repondent-provided value in the SIPP. Date of birth was selected from the first non-missing value in the following files: (i) SSA's Master Benefits Record (MBR) file, (ii) the Census Bureau's Person Characteristic File (PCF) whose main input is the SSA Numident file, and (iii) SSA's Supplemental Security Record (SSR) file. Thus, this variable is administrative and sometimes differs from the birth date reported in the SIPP survey itself. When missing due to the lack of a validated SSN for the SIPP respondent, date of birth was imputed using date of birth from the Census-internal version of the SIPP.
We chose the administrative source for two reasons. First, the administrative birth date was more often consistent with the other administrative data (benefits and earnings). For example, when age was calculated using the administrative birth date, there were fewer individuals who appeared to retire before age 62. Second, the differences between the administrative birth date and the birth date reported in the survey helped to increase the difficulty of re-identifying a record in

the original SIPP public use data from a record in the synthetic data, thus improving the confidentiality protections. This variable is coded as a SAS date variable. This format gives the number of days between the date of birth and January 1, 1960. An individual born on January 1, 1959 would have birthdate=-365 and an individual born on January 1, 1961 would have birthdate=365.
In addition to help researchers we have created the variables:
brthdy_final
brthmn_final
brthyr_final
These variables are numeric, length 3 and contain the day, month, and year respectively of the birthdate.

---

flag_deathdate_exist — Flag: Existence of Date of Death

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Death date does not exist, respondent did not die during this interval
1	Death date exists, respondent died during this interval

Flag to indicate that this respondent died after being interviewed and by the end of administrative time series.

---

deathdate — Date of Death

---

type: numeric (double)

range: positive values

Date of death from administrative data. This variable also is obtained using a hierarchy of administrative sources: (i) SSA's MBR file, (ii) the Census PCF with death information coming from the SSA Numident and Master Death Files, and (iii) SSA's SSR file. This variable is coded as a SAS date variable.
In addition to help researchers we include:
deathdy_final
deathmn_final
deathyr_final
which are numeric 3 and contain the day, month, and year, respectively, of the date of death.

MBR/PHUS Variables

The Master Benefits Records (MBR) is SSA's main file to track who is receiving Old Age Survivor and Disability (OASDI) benefits, the reason for receipt, and the monthly benefit amounts payable to the individual.

The Payment History Update System (PHUS) contains actual payments delivered to OASDI beneficiaries. The data from the PHUS may differ from what are contained on the MBR due to discrepancies between the timing of SSA awarded amounts and the actual payments made to participants. This situation would be expected to affect disability cases more than aged cases because it takes more time to establish eligibility to receive disability. Individuals are eligible to receive benefits due to their own earnings history and age, as well as due to a spouse's earnings history and age. In this section retirement and disability are "own" benefits while aged spouse, widowed spouse, and other are "spouse" benefits. The age requirements for receiving each type of benefit are as follows:

Retire	- minimum age 62 (reduced benefit), full retirement age (full benefit)
Disability	- under age 65 or full retirement age, whichever is greater; at full retirement age, these benefits convert to retirement.
Aged Spouse	- minimum age 62 (reduced benefit), full retirement age (full benefit), spouse must be retired or disabled
Widowed Spouse	- minimum age 60 (reduced benefit), full retirement age (full benefit), spouse must be deceased
Other	- no age requirements

Until the year 2000, the full retirement age was 65. From 2000 to 2022, the full retirement age is increasing by 2 months each year so that by 2022 the full retirement age will be 67. The benefits reported in this section are total benefits received at a point in time. The MBR research extract provided by SSA to create the Gold Standard contains information about different reasons for receiving benefits but does not always allow the amount due to each reason to be accurately separated from the total. Hence we have elected to report total benefits at a point in time and researchers should be careful to note that when an individual is receiving both own retirement and aged spouse benefits, the amounts listed for each benefit type will be redundant, i.e. there is really only one total amount and two reasons for receiving it. SSA calculates benefits based on an individual's lifetime earnings history following rules which they publish in "Annual Statistical Supplement to the Social Security Bulletin", available for each tax year on the Social Security website, www.ssa.gov.

---

flag_in_mbr — Flag: in MBR

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Respondent was not matched to MBR
1	Respondent was matched to MBR

This flag indicates that a person matched to the SSA Master Beneficiary File (MBR). The person's SSN showed up in the MBR because they received benefits of some kind.

---

mbr_agedsp — MBR: receive agedsp benefit

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Does not receive monthly agedsp benefit
1	Receives monthly agedsp benefit

Indicates that individual received aged spouse benefits at some point during the time period covered by the MBR extract. This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

---

mbr_agedsp_stdate — MBR: startdate of benefit

---

type: numeric (double)

range: positive/negative values

Date when the person first began receiving aged spouse benefits, conditional on having ever received this type of benefit.

---

mbr_agedsp_totamt — MBR: total monthly benefit

---

type: numeric (double)

range: positive values

Total monthly amount of benefits received at beginning of aged spouse benefit entitlement. In most cases this amount is from the same month as in MBR_agedsp_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

---

phus_agedsp_stdate — PHUS: startdate of benefit

---

type: numeric (double)

range: positive values

Date aged spouse benefits began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR aged spouse benefit start date. It also must be 1984 or later because PHUS data began in 1984.

---

phus_agedsp_totamt — PHUS: total monthly benefit

---

type: numeric (double)

range: positive values

Total amount of benefits as recorded in the PHUS in the first month of receiving aged spouse benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

---

mbr_disab — MBR: receive disab benefit

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Does not receive monthly disability benefit
1	Receives monthly disability benefit

Indicates that individual received disability benefits at some point during the time period covered by the MBR extract. This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

---

mbr_disab_stdate — MBR: startdate of benefit

---

type: numeric (double)

range: positive/negative values

Date at which individual began receiving own disability benefits. This date must be before individual reaches the full retirement age (FRA). FRA depends on the year the person reaches age 62. Any individual who turned before 2000 had FRA=65 years old. Beginning in 2000, any individual turning 62 had full retirement age of 65 years + 2*(year_age_62 - 1999) months.

---

mbr_disab_totamt — MBR: total monthly benefit

---

type: numeric (double)

range: positive values

Total monthly amount of benefits received at beginning of disability benefit entitlement. In most cases this amount is from the same month as in MBR_disab_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reason (i.e. dual entitlement).

---

phus_disab_stdate — PHUS: startdate of benefit

---

type: numeric (double)

range: positive values

Date disability benefits began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR disability benefit start date. It also must be 1984 or later because PHUS data began in 1984.

---

phus_disab_totamt — PHUS: total monthly benefit

---

type: numeric (double)

range: positive values

Total amount of benefits as recorded in the PHUS in the first month of receiving own disability benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

---

mbr_other — MBR: receive other benefit

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Does not receive other monthly benefits
1	Receives other monthly benefits

Indicates that individual received other benefits at some point during the time period covered by the MBR extract. There were four types of other benefits: Spouse caring for minor children (TOB=4), Widow(er) caring for minor children (TOB=6), Disabled Widow(er) (TOB=7), Adult disabled in childhood (TOB=8). These types were combined because they are relatively rare and would be a confidentiality risk if reported on an individual basis. It is important to note that all benefits included in our 5 benefit types are benefits that are received by adults. We do not include information about payments received as a child.
This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

---

mbr_other_stdate — MBR: startdate of benefit

---

type: numeric (double)

range: positive/negative values

Date when the person first began receiving other benefits, conditional on having ever received this type of benefit.

---

mbr_other_totamt — MBR: total monthly benefit

---

type: numeric (double)

range: positive values

Total monthly amount of benefits received at beginning of other benefit entitlement. In most cases this amount is from the same month as in MBR_other_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

---

phus_other_stdate PHUS: startdate of benefit

---

type: numeric (double)

range: positive values

Date other benefit began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR other benefit start date. It also must be 1984 or later because PHUS data began in 1984.

---

phus_other_totamt PHUS: total monthly benefit

---

type: numeric (double)

range: positive values

Total amount of benefits as recorded in the PHUS in the first month of receiving other benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

---

mbr_retire MBR: receive retire benefit

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Does not receive monthly retire benefit
1	Receives monthly retire benefit

This variable indicates that a person received retirement benefits at some point during the time period covered by the MBR extract (for dates see ...). These benefits were the result of the individual's own earnings history.

---

mbr_retire_stdate MBR: startdate of benefit

---

type: numeric (double)

range: positive/negative values

Date when the person first began receiving own retirement benefits, conditional on having ever received this type of benefit.

---

mbr_retire_totamt — MBR: total monthly benefit

---

type: numeric (double)

range: positive values

Total monthly amount of benefits received at beginning of own retirement benefit entitlement. In most cases this amount is from the same month as in MBR_retire_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

---

phus_retire_stdate — PHUS: startdate of benefit

---

type: numeric (double)

range: positive values

Date retirement benefits began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR retirement benefit start date. It also must be 1984 or later because PHUS data began in 1984.

---

phus_retire_totamt — PHUS: total monthly benefit

---

type: numeric (double)

range: positive values

Total amount of benefits as recorded in the PHUS in the first month of receiving own retirement benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

---

mbr_widowsp — MBR: receive widowsp benefit

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	Does not receive monthly widowsp benefit
1	Receives monthly widowsp benefit

Indicates that individual received widowed spouse benefits at some point during the time period covered by the MBR.

This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

---

mbr_widowsp_stdate — MBR: startdate of benefit

---

type: numeric (double)

range: positive/negative values

Date when the person first began receiving widowed spouse benefits, conditional on having ever received this type of benefit.

---

mbr_widowsp_totamt — MBR: total monthly benefit

---

type: numeric (double)

range: positive values

Total monthly amount of benefits received at beginning of widowed spouse benefit entitlement. In most cases this amount is from the same month as in MBR_agedsp_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

---

phus_widowsp_stdate — PHUS: startdate of benefit

---

type: numeric (double)

range: positive values

Date widowed spouse benefits began being paid, as recorde in the PHUS. This date must be greater than or equal to the MBR widowed spouse benefit start date. It also must be 1984 or later because PHUS data began in 1984.

---

phus_widowsp_totamt — PHUS: total monthly benefit

---

type: numeric (double)

range: positive values

Total amount of benefits as recorded in the PHUS in the first month of receiving widowed spouse benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

Supplemental Security Record Variables

The Supplemental Security Records (SSR) is SSA's main file to track who is receiving Supplemental Security Income (SSI) benefits and the monthly benefit amounts payable. SSI benefits are paid to elderly, blind, or disabled individuals who fall below certain income threshholds. Eligibility and federal payment standards are uniform across all states but states have the option to supplement federal payments. The payment included here is the total of both federal and state SSI payments.

---

flag_in_ssr — Flag: in SSR

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	SSN not found in SSA Supplemental Security Records (SSR)
1	SSN found in SSR

This flag indicates that a person's SSN was found in the SSA Supplemental Security Records (SSR). This database tracks people who received SSI.

---

ssr_ssi_date_initial_entitle — SSR: SSI Date of Initial Entitlement

---

type: numeric (double)

range: positive values

Date of initial entitlement to SSI benefits.

---

ssr_ssi_amt_initial — SSR: SSI Amount - Initial ($2000)

---

type: numeric (double)

range: positive values

Amount of monthly SSI payment at time of initial receipt.

IRS/SSA Variables

The Census Bureau sent a list of validated SSNs from the seven included SIPP panels to SSA and extracts from the Master Earnings File (Summary and Detailed Earnings Records), Master Beneficiary Record, Supplemental Security Record, 831 Disability File, and Payment History Update System were created. The variables from these files that are included in the SSB are described below.

Not all SIPP respondents have linkages to SSA/IRS administrative data, including: those who refused to provide their SSN; those whose SSNs were not validated; and those with valid SSNs who never worked, and never applied for

benefits or received benefits. In the Gold Standard, individuals without a validated SSN or without SSA/IRS administrative records had missing data for all SSA/IRS-derived variables described below. Among these people, those respondents without a validated SSN had all administrative data imputed as part of the data completion process. Hence in the completed Gold Standard and the synthetic data, only individuals with no work or benefit history have zero earnings and missing benefits.

---

flag_valid_ssn — Flag: Valid SSN

---

type: numeric (double)

range: [0,1]

Numeric	Label
0	No validated SSN and hence no link to administrative data
1	Has validated SSN to link to administrative data

IRS/SSA Variables

The Census Bureau sent a list of validated SSNs from the seven included SIPP panels to SSA and extracts from the Master Earnings File (Summary and Detailed Earnings Records), Master Beneficiary Record, Supplemental Security Record, 831 Disability File, and Payment History Update System were created. The variables from these files that are included in the SSB are described below.

Not all SIPP respondents have linkages to SSA/IRS administrative data, including: those who refused to provide their SSN; those whose SSNs were not validated; and those with valid SSNs who never worked, and never applied for benefits or received benefits. In the Gold Standard, individuals without a validated SSN or without SSA/IRS administrative records had missing data for all SSA/IRS-derived variables described below. Among these people, those respondents without a validated SSN had all administrative data imputed as part of the data completion process. Hence in the completed Gold Standard and the synthetic data, only individuals with no work or benefit history have zero earnings and missing benefits.

Detailed Earnings Record Variables

The Detailed Earnings Records (DER) contains historical earnings reports for each person and job held from 1978 onwards. These reports include self employment income. Earnings are not capped at the taxable maximum. For each tax year, we summed DER information for each person across all jobs and self employment to create a total earnings amount.

---

defer_der_fica — DER: Deferred FICA

---

type: numeric (double)

range: positive values

Deferred earnings from jobs covered by FICA tax; summed across all employers in the DER to give a person-level total for each year. While the variable exists on the Gold Standard for the years 1978-1986, it is always missing in this time period. The year 1987 is the first year with positive deferred wages. On the synthetic and completed gold standard files, we only keep 1990-2006 because so few people had deferred wages between 1987 and 1989 that we could not reliably synthesize these variables.

---

nondefer_der_fica DER: Non-Deferred FICA

---

type: numeric (double)

range: positive values

Non-deferred earnings (i.e. paid to individual) from jobs covered by FICA tax; summed across all employers in the DER to give a person-level total for each year.

---

defer_der_nonfica DER: Deferred Non-FICA

---

type: numeric (double)

range: positive values

Deferred earnings from jobs NOT covered by FICA tax; summed across all employers in the DER to give a person-level total for each year. While the variable exists on the Gold Standard for the years 1978-1986, it is always missing in this time period. The year 1987 is the first year with positive deferred wages. On the synthetic and completed gold standard files, we only keep 1990-2008 because so few people had deferred wages between 1987 and 1989 that we could not reliably synthesize these variables.

---

nondefer_der_nonfica DER: Non-Deferred Non-FICA

---

type: numeric (double)

range: positive values

Non-deferred earnings (i.e. paid to individual) from jobs NOT covered by FICA tax; summed across all employers in the DER to give a person-level total for each year.

---------- Summary Earnings Record Variables ----------

The SSA/IRS Summary Earnings Records (SER) contain historical person-level earnings data. In addition to an array of annual FICA-taxed earnings

(1951-2006) that are capped at the FICA taxable maximum, the SER provides information regarding quarters of covered work. Quarters of covered work are utilized by SSA to determine eligibility for participation in its old age, survivors, and disability insurance (OASDI) programs.

---

totearn_ser — SER: Total Earnings

---

type: numeric (double)

range: positive values

Annual earnings taxed by FICA; these variables include earnings only up to the FICA taxable maximum, i.e., these earnings measures are capped.

---

wqc_yrtot — SER: Annual Total Covered Quarters of Work

---

type: numeric (double)

range: [0,4]

Indicates the total number of quarters of FICA-covered work.

SIPP Arrays

This section contains variables that are repeated over time, reflecting the panel nature of the SIPP. For these time series variables, individuals will only have values for the months and years in which they participated in a SIPP panel. We also include only years fully covered by one of the SIPP panels. Hence there are no data for 1995 or 2000 since neither of these years were fully covered by a panel.

---

totearn — Total Earnings

---

type: numeric (double)

range: positive values

This is recode variable totearn from the public use file, just without the top codes.

---

tothours — Total Hours Worked at All Jobs

---

type: numeric (double)

range: positive values

Total number of hours worked at all jobs in a given month.

---

hicov Health Insurance Coverage

type: numeric (double)

range: [0,1]

Numeric	Label
0	Respondent did not have health insurance coverage during this month
1	Respondent had health insurance coverage during this month

hiemp Health Insurance Coverage from Employer

type: numeric (double)

range: [0,1]

Numeric	Label
0	Respondent did not have employer-provided health insurance
1	Respondent had employer-provided health insurance

wksjob Weeks at a Job

type: numeric (double)

range: [0,5]

Total number of weeks worked at a job in a given month.

wkswp Weeks With Pay

type: numeric (double)

range: [0,5]

Total number of weeks worked with pay in a given month.

totinc Total Personal Income

type: numeric (double)

range: positive/negative values

Personal income summed from all sources.

Geographic Variables

These variables are included for internal use purposes and describe the geography of the address where the SIPP household was first interviewed. These variables are panel specific, with the year subscript at the end of the variable name representing the source panel. Individuals will have nonmissing geography information only for the variables that correspond to their SIPP panels.

---

state — State of Residence

---

type: numeric (double)

range: [1,63]

Numeric	Label
1	Alabama
2	Alaska*
4	Arizona
5	Arkansas
6	California
8	Colorado
9	Connecticut
10	Delaware
11	DC
12	Florida
13	Georgia
15	Hawaii
16	Idaho*
17	Illinois
18	Indiana
19	Iowa*
20	Kansas
21	Kentucky
22	Louisiana
23	Maine*
24	Maryland
25	Massachusetts
26	Michigan
27	Minnesota
28	Mississippi*
29	Missouri
30	Montana*
31	Nebraska
32	Nevada
33	New Hampshire
34	New Jersey
35	New Mexico*
36	New York
37	North Carolina
38	North Dakota*
39	Ohio
40	Oklahoma
41	Oregon
42	Pennsylvania
44	Rhode Island

45 South Carolina
46 South Dakota*
47 Tennessee
48 Texas
49 Utah
50 Vermont*
51 Virginia
53 Washington
54 West Virginia*
55 Wisconsin
56 Wyoming*
61 *see description
62 *see description
63 *see description

FIPS State Code for state of residence first recorded in the SIPP. For married couples, we take the state value for both partners at the same point in the survey when we first observed the marriage. For individuals who never have an observed marriage during the panel, we take their first ever reported state value.
*All panels prior to 2004 group some states together and give only one code for the group.
In the 1984 panel, the following state groupings are used:
90=Idaho, New Mexico, South Dakota, Wyoming
91=Mississippi, West Virginia
In the 1990, 1991, 1992, and 1993 panels, the following state groupings are used:
61=Maine, Vermont
62=Iowa, North Dakota, South Dakota
63=Alaska, Idaho, Montana, Wyoming
In the 1996 and 2001 panels, the following state groupings are used:
61=Maine, Vermont
62=North Dakota, South Dakota, Wyoming
For these panels, the individual FIPS code will not appear for states contained in a group.

.