INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
HEALTH AND RETIREMENT STUDY
UNIVERSITY OF MICHIGAN

# Health and Retirement Study
# Imputations of Lifetime Earnings Records

## Data Description and Technical Documentation

Chichun Fang
Institute for Social Research
University of Michigan

Version 1.0
February 2018

# Contents

# 1 Overview

The *Imputations of Lifetime Earnings Records* is a restricted data release based on earnings records of respondents in the Health and Retirement Study (HRS) provided by the Social Security Administration (SSA).

The HRS is a national longitudinal study of the economic, health, marital, family status, and public and private support systems of older Americans, and is a rich data source for researchers and policymakers who study aging. The National Institute on Aging provided funding (NIA U01 AG009740), with supplemental support from the SSA. The study is conducted by the Institute for Social Research at the University of Michigan.

The HRS gratefully acknowledges the special assistance of the SSA's Office of Research and Statistics for their assistance in retrieving the administrative records of HRS respondents who gave consent for those records to be used for research purposes.

This restricted data set is intended for exclusive use by you and the persons specified in the *Agreement for Use of Restricted Data from the Health and Retirement Study* and/or the *Supplemental Agreement with Research Staff for Use of Restricted Data from the Health and Retirement Study*.

# 2 Construction of the Imputed Earnings Records Data

The main purpose of this data product is to provide researchers a set of earnings records in which we replace the top-coded or masked numbers in SSA's earnings records with imputed ones. We further split a respondent's earning into three components: earnings in covered[1] employment, earnings in non-covered employment, and earnings in OASDI-covered self-employment. Our ability to construct and/or impute these variables is constrained by the nature and granularity of SSA earnings records, which changed over time. We describe how these three components of earnings were constructed/imputed in each period in a reverse chronological order, as more records were masked or top-coded in earlier years.

[1] In this document, the terms "covered", "covered by Social Security", and "OASDI-covered" are used interchangeably.

To understand the procedures detailed in this section, the user needs to know the type(s) of SSA earnings data available to the HRS. For most respondents, summary records were available between 1951 and 1977, and detailed records were available from 1978 onward. The following reverse chronological order pertains to these respondents. However, some respondents from the older cohorts only consented when they initially entered in the HRS in 1992 or 1993 but never re-consented ever since. The HRS only has summary records but no detailed records for them, and the records stopped in 1991 or 1992. For these respondents, the earnings were always imputed using the procedure as in "between 1951 and 1977" subsection below, regardless the corresponding year of the earnings record.

### 2.1 Earnings from 1994 Onward

In this period, all available earnings records are detailed records. In a record that reported positive wage earnings, we can tell whether the job was from covered or non-covered employment by examining whether the earning was OASDI taxable. If the record was from self-employment, the earnings would be reported as self-employment earnings subject to OASDI tax and Medicare tax (two separate variables). Accordingly, we can categorize earnings as covered earnings, non-covered earnings, or self-employment earnings.

For covered or non-covered employment, wage earnings in detailed earnings records were not top-coded. For self-employment, earnings subject to OASDI tax were top-coded at OASDI taxable maximum, and earnings subject to Medicare tax were top-coded at Medicare taxable maximum. However, Medicare taxable maximum had been suspended since 1994, and every dollar was subject to Medicare tax. Consequently, self-employment earnings from 1994 onward could be treated as not being top-coded as well.

Both covered and non-covered earnings were operationally defined as the maximum of (a) and (b), where:

- (a) was the sum of two components: (a1) earnings and tips subject to federal income tax, and (a2) contributions to qualified plans that were income tax-deferred;

- (b) was earnings and tips subject to Medicare tax.

In most cases, (a) and (b) should be equal. Self-employment earnings were simply the earnings subject to Medicare tax.

However, in order to preserve confidentiality, earnings in SSA records were "masked" if they were above $250,000. An amount between $250,000 and $300,000 would show up as ".x" in the detailed records, an amount between $300,000 and $500,000 would show up as ".y", and amount greater than $500,000 would show up as ".z". We first imputed these masked numbers using the parametric assumption in 4.1 and the method in 5.1, assuming the earnings above twice of the OASDI taxable maximum in the corresponding year followed Pareto distribution. We then took the sum of covered earnings, non-covered earnings, and self-employment earnings (using the imputed numbers if the original ones were masked) at the yearly level. If a respondent did not have any earnings, and hence no earnings record at all, in a given year, all three earnings variables were defined as missing.

### 2.2 Earnings between 1991 and 1993

This period was the same as 1994 onward with one exception: the Medicare taxable maximum was in effect. Top-coding at Medicare taxable maximum affected self-employment earnings but did not affect earnings in covered or non-covered employment. We used the same strategy as previously explained to impute for the earnings records masked as ".x", ".y", or ".z". That gave us un-top-coded, un-masked earnings in covered and non-covered employment. We then used the nearest neighbor matching method in Section 5.2 to impute the self-employment earnings that were top-coded.

### 2.3 Earnings between 1978 and 1990

Between 1978 and 1990, employment earnings from detailed records were still not top-coded, but self-employment earnings were top-coded at the OASDI taxable maximum. We used the same Pareto interpolation to impute the numbers were masked, and then the same nearest neighbor matching method as in 1991-1993 to impute the earnings of self-employed respondents whose earn-

ings were top-coded.

### 2.4 Earnings between 1951 and 1977

Only summary records were available between 1951 and 1977. As previously mentioned, for some respondents in the older cohorts who only consented in 1992 or 1993 but never re-consented later, their earnings between 1978 and 1991/1992 were also imputed this way.

Summary earnings records differed from detailed earnings records in the following aspects: (1) only earnings from covered employment or covered self-employment were available; (2) in each year, all covered earnings were lumped together and reported as one single number; and (3) the reported covered earnings were top-coded at the OASDI taxable maximum. Due to such nature, non-covered earnings were not available during the years when we only had summary earnings records. For this period, we did not attempt to impute non-covered earnings; neither did we attempt to separate earnings from covered employment and covered self-employment. All the earnings in these years would be categorized under "covered earnings" during this period, earnings from non-covered employment as well as earnings from covered self-employment were both set to ".n". This allowed the researchers to differentiate years in which no earnings records were available vs. years in which non-covered and self-employed earnings were not available. See Section 7 for details.

Except for the respondents who only consented in 1992 or 1993 but never re-consented ever since, a "pattern of coverage" variable was available in the summary records. With some assumptions regarding earnings patterns, this variable told us which quarter in a calendar year the taxable maximum was hit. This variable and the assumption we made are explained in Section 3.

For the top-coded records in respondent-year that we knew when the taxable maximum had been hit (through the 'pattern of coverage" variable), we used Pareto or Log-Normal interpolations to impute the earnings. The parametric assumptions and procedures of imputation are detailed in Sections 4.1, 4.2, and 5.1. That left us the earnings of respondents whose pattern of coverage information was not available or who were self-employed (self-employed earnings were reported to SSA annually rather than quarterly). Covered earnings of these respondents were imputed using nearest neighbor matching as Section 5.2.

# 3 Pattern of Coverage in Summary Earnings Records

## 3.1 Pattern of Coverage Variable

Between 1951 and 1977, employers were required to report quarterly earnings information for each employee to SSA. A quarter of coverage (QC) was credited for each quarter in which an employee received $50 or more in covered earnings and tips (Olsen and Hudson, 2009), up to four QCs in each year. An employee who reached the OASDI taxable earnings maximum would get 4 QCs; after the taxable maximum was hit, the employer stopped reporting earnings information to SSA for this employee for the rest of the year.

For each quarter that more than $50 was reported, an individual would receive a "1" in the pattern of coverage variable, and "0" otherwise. If earnings information was reported in each quarter, the coverage pattern would be "1111." However, if earnings information was only reported for the first two quarters, the coverage pattern would be "1100." Among the individuals who received 4 QCs and whose earnings were reported as the taxable maximum, this variable essentially indicated when the taxable maximum was hit—because the employer was no longer required to report earnings information for this employee for the rest of the year.

Pattern of coverage information was available for most HRS respondents between 1951 and 1997. The only exceptions were those among the earliest cohorts who consented to give the HRS their Social Security earnings records only in their first interview in 1992 or 1993 but never re-consented in later waves. These consent forms did not include detailed earnings records, and neither did they include the pattern of coverage variable. The pattern of coverage was "9999", indicating "not applicable" for each respondent in this group between 1951 and 1991. We explained how alternative imputation strategies were developed for this group in Section 5.2.

There is one more exception to this quarter of earnings coverage variable. Self-employment earnings and farm wages were reported annually to SSA, rather than quarterly. For the purpose of determining the amount of covered earnings, SSA counted earnings information reported by employers first, and then self-employment earnings (Olsen and Hudson, 2009). Hence, there can be three possible scenarios among individuals with self-employment earnings who hit the OASDI

taxable maximum:

- Earnings reported by the employer already reached the taxable maximum. The pattern of coverage information would only rely on employee report, as if there were no self-employment earnings. Total QC was 4.
- Earnings reported by the employer plus self-employment earnings reached the taxable maximum. The pattern of coverage information would only rely on employee report. The individual would receive QC credits from both employment and self-employment earnings, but the total QC would still be capped at 4.
- There were no earnings reported by any employer, and self-employment earnings reached the taxable maximum. The pattern of coverage would be "0000." Total QC was 4.

The first two scenarios did not affect our imputation. We considered the pattern "0000" as pattern of coverage not available and treated them the same as "9999" in Section 5.2.

### 3.2 Assumptions about Earnings Pattern

The OASDI taxable maximum is $Y_t^*$ in year $t$ (between 1951 and 1977). $Y_{it}$ is the earnings of individual $i$ in year $t$. Hence, in the data we observe:

$$Y_{it} = \begin{cases} Y_{it} & \text{if } Y_{it} \leq Y_t^* \\ Y_t^* & \text{if } Y_{it} > Y_t^* \end{cases}$$

Among those who hit the taxable cap, we knew whether they hit the cap in Q1 (the first quarter in the calendar year), Q2, Q3, or Q4. We made two assumptions about the earnings of those who hit the taxable cap $Y^*$ (the subscript $t$ is omitted for simplicity):

- Annual earnings were assumed to distribute uniformly across the four calendar quarters.
- Those who hit the cap in Q1 had higher annual earnings than those who hit the cap in Q2 than those who hit the cap in Q3, etc.

These two assumptions essentially determined the range of possible annual earnings relative to $Y^*$ given the calendar quarter when the cap was hit. For example, if a respondent hit the cap in Q2,

her annual earnings in that year must be between 200% and 400% of $Y^*$. She would have hit the cap in Q1 had her earnings been equal to or higher than 400% of $Y^*$, and she would not have hit the cap until Q3 had her earnings been lower than 200% of $Y^*$. Similarly, those who hit the cap in Q3 would have earnings between 133% and 200% of $Y^*$, and those who hit the cap in Q4 would have earnings between 100% and 133% of $Y^*$.

# 4 Parametric Assumptions about the Earnings Distribution

Consistent with the earnings distribution literature, we assumed that the earnings distribution had two "segments." In any given year, the earnings distribution was log-normal up until a certain threshold and became Pareto beyond that threshold. We defined such threshold as 200% of $Y^*$. In other words, in any given year, those who hit $Y^*$ in Q1 or Q2 had earnings following a Pareto distribution, and the rest of sample had earnings following a log-normal distribution.[2] Such definition of the threshold was based on the actual earnings patterns observed in the data. Between 1951 and 1977, about 1% of respondents hit $Y^*$ in Q1 or Q2. There were some variations during the span. The proportion of respondent hitting $Y^*$ in Q1 or Q2 increased from less than 0.5% in the early 1950s, peaked at 2.6% in 1965, and then decreased to less than 1% in 1977.

## 4.1 Pareto Distribution

In a Pareto distribution, the probability of earnings $Y$ larger than a given level $X$ is determined by two parameters, $K$ and $\alpha$:

$$Prob(Y > X) = (K/X)^{\alpha} \tag{1}$$

$K > 0$ is the minimum level of earnings to which the Pareto distribution applies, and $\alpha$ determines the shape of the distribution.

Let $F_{Q1}$ be the proportion of respondents who did *not* hit the taxable cap $Y^*$ by Q1, and $F_{Q2}$ be the proportion of respondents who did *not* hit $Y^*$ by Q2. By our assumption, those who hit $Y^*$ by Q1 had earnings at least 400% of $Y^*$, and those who hit $Y^*$ by Q2 (i.e. in Q1 or Q2) had earnings

[2] Kopczuk, Saez, and Song (2010) also used SSA earnings records, and they only assumed the earnings of those who hit $Y^*$ in Q1 to be Pareto. We were not able to do so as we did not have the same quarterly earnings data that they had.

at least 200% of $Y^*$. Hence, $F_{Q2}$ and $F_{Q1}$ can be seen as cumulative densities of respondent whose earnings were below 200% of $Y^*$ and 400% of $Y^*$, respectively. Equation (1) implies

$$\begin{cases} 1 - F_{Q1} = (K/4Y^*)^\alpha \\ 1 - F_{Q2} = (K/2Y^*)^\alpha \end{cases}$$

Solving the equations yields $\hat{\alpha} = \ln[(1 - F_{Q1})/(1 - F_{Q2})]/\ln[0.5]$ and $\hat{K} = 4Y^* \cdot (1 - F_{Q1})^{(1/\hat{\alpha})}$. We allowed $\hat{\alpha}$ and $\hat{K}$ to vary over time and estimated a pair of $(\hat{\alpha}, \hat{K})$ for each year between 1951 and 1977. Our $\hat{\alpha}$ and $\hat{K}$ were generally higher than those reported in Feenberg and Poterba (1993) but comparable to those reported in Picketty and Saez (2003).[3]

Let $Y_i > \hat{K}$ be the earnings of respondent $i$, and $F_i$ is the cumulative density of respondents below $Y_i$. Substituting $\hat{\alpha}$ and $\hat{K}$ in equation (1) yields

$$\begin{aligned} & 1 - F_i = (\hat{K}/Y_i)^{\hat{\alpha}} \\ \iff \quad & Y_i = e^{\ln \hat{K}^* - \frac{1}{\hat{\alpha}} \ln(1 - F_i)} \end{aligned} \tag{2}$$

That is, for any given $F_i$, the corresponding $Y_i$ can be calculated accordingly using equation (2). Since $100 * F_i$ numerically corresponds to the percentile of $Y_i$ in the earnings distribution, this means we can calculate $Y_i$ as long as we know where respondent $i$ is in the earnings distribution.

### 4.2 Log-Normal Distribution

For respondents who hit $Y^*$ in Q3 (earnings between 133% and 200% of $Y^*$) or Q4 (earnings between 100% and 133% of $Y^*$), we assumed the earnings were distributed as log-Normal.

Let $F_{Q3}$ be the proportion of respondents who did *not* hit the taxable cap $Y^*$ by Q3, and $F_{Q4}$ be the proportion of respondents who did *not* hit $Y^*$ by Q4. $X_{Q3}$ and $X_{Q4}$ are drawn from a standard normal distribution such that $\Phi(X_{Q3}) = F_{Q3}$ and $\Phi(X_{Q4}) = F_{Q4}$, where $\Phi(\cdot)$ is the cumulative density function of a standard normal distribution. Also, $Y_{Q3} = 1.33Y^*$ and $Y_{Q4} = Y^*$ are the minimum earnings required in order to hit the taxable cap by Q3 and Q4, respectively. By the

[3]Both Feenberg and Poterba (1993) and Picketty and Saez (2003) used data that were representative of all U.S. taxpayers. However, our data were not supposed to be representative of all U.S. taxpayers between 1951 and 1977 (or at any point of time for that matter) due to the nature of the HRS. Hence, we did the comparison to assess how the distributions compared rather than to check the validity of our estimates.

nature of a log-Normal distribution, we have

$$\begin{cases} \ln(Y_{Q3}) = \mu + X_{Q3} \cdot \sigma \\ ln(Y_{Q4}) = \mu + X_{Q4} \cdot \sigma \end{cases}$$

where $\mu$ and $\sigma$ are the parameters that determine the log-Normal distribution. Solving the equations yields $\hat{\sigma} = [\ln(Y_{Q4}) - \ln(Y_{Q3})]/[X_{Q4} - X_{Q3}]$ and $\hat{\mu} = ln(Y_{Q3}) - X_{Q3} \cdot \hat{\sigma}$. We also allowed $\hat{\mu}$ and $\hat{\sigma}$ to vary over time, and they were calculated with $Y_{Q3}$ and $Y_{Q4}$ being replaced by year-specific $1.33Y_t^*$ and $Y_t^*$, respectively. With $\hat{\mu}$ and $\hat{\sigma}$, earnings at any given percentile within the range can be calculated.

# 5 Imputation Strategy

## 5.1 Pareto and Log-Normal Interpolations

With the parametric assumptions specified in Section 4, the interpolation can be calculated with one single piece of information: where the respondent was in the earnings distribution.

In years that detailed earnings records were available, we counted the numbers of respondent/record with un-masked earnings as well as respondent/record masked as ".x", ".y", or ".z". For example, say 980 records were not masked, 10 were ".x", 7 were ".y", and 3 were ".z". We could infer that those masked as ".x" must be between the $98^{th}$ and $99^{th}$ percentiles of earnings distribution; ".y" must be between the $99^{th}$ and $99.7^{th}$ percentiles, and "z" must be above the $99.7^{th}$ percentile. In years that only summary earnings records were available, assumptions in Section 3.2 facilitated similar calculations. This is the first step of our imputation, which bracketed the "range" of earnings percentiles each individual belonged to in a given year. If there were no autocorrelation in earnings over time, we could just randomly draw a number from the corresponding range of percentiles for each individual in each year, and plug the drawn percentile back to the corresponding earnings distribution parameters to obtain the imputed earnings.

The autocorrelations in earnings were modeled in this data product as autocorrelation in the rank of earnings, rather than the level of earnings. Conceptually, our algorithm can be explained with a simplified example. Assuming that we only have two years of earnings records; 1% of records were top-coded in year 1 and 10% were top-coded in year 2. Using the information in year 1, we

could calculate where everybody was in the ranking using the non-top-coded earnings up to the $99^{th}$ percentile. We would also know who were among the top 1%, although we did not know exactly where they were along the distribution. If we assumed the relative ranking stayed the same in both years, these two pieces of information allowed us to know the ranking of each respondent was up to the $99^th$ percentile as well as who were among the top 1%. Consequently, although 10% respondents had earnings top-coded in year 2, we could still infer where they were along the distribution and impute the earnings accordingly.

In this data product, we relaxed the assumption that relative ranking stayed the same over time. We instead assumed that rankings were "smooth" and predicted individual ranking with a moving average. The ranking was done by cohort, to account for the correlation between earnings rank and age. There were also enough granularities in the data to further allocate respondents into various categories of top-coding: Q1 through Q4 between 1951 and 1977, and ".x" through ".z" 1978 onward. This is the second step of our imputation.

Finally, we ranked the predicted ranking from Step 2 within the "bracket" of ranking obtained in Step 1. That is, among the 10 respondents categorized between the $98^{th}$ and $99^{th}$ percentiles in Step 1, the respondent with lowest predicted ranking from Step 2 would be assigned as the $98^{th}$ percentile, the respondent with the second lowest predicted ranking be the $98.1^{th}$ percentile, ..., and the respondent with the highest predicted ranking be the $98.9^{th}$ percentile. Such information was then plugged back to the earnings distributions parameters to obtain the imputed earnings.

This imputation is hence akin to the imputation of responses in unfolding brackets in the HRS core survey. We preserved the "bracket information" (the range of earnings distribution where the respondent belonged to) and supplemented it with earnings ranking from adjacent years.

### 5.2 Nearest Neighbor Matching

The intuition behind this matching is to match a respondent $i$ (whose earnings could not be imputed using the previous method) with an otherwise similar respondent $j$ (whose above-the-cap earning was either imputed using the previous method or directly available in the detailed earnings records), and then assign respondent $j$'s imputed or observed earning to respondent $i$.

To do so, we first estimated a prediction equation with all the respondents whose (a) earnings were below the cap, (b) earnings were top-coded at the cap in the summary records but imputed using the method in Section 5.1, or (c) earnings were at or above the cap but directly available in the detailed records. The prediction equation was estimated separately by year and HRS cohort.

The left-hand side of the prediction equation was the logarithm of earnings in year $t$.[4] The right-hand side included the cumulative covered earnings between 1951 and year $t$ (price-adjusted using the Social Security wage index)[5], number of years the respondent hit the taxable cap between 1951 and year $t$, total quarters of earnings between 1951 and year $t$, and birth year dummy variables. As previously mentioned, the prediction equation was estimated separately by year and HRS cohort. The coefficients from the prediction equation were used to calculate a "predicted earning" for everybody, pattern of coverage information available or not.

Each respondent whose earning was still top-coded at this stage was matched to a respondent (a) whose earnings was above the taxable cap and had been imputed previously in Section 5.1 or directly observed, and (b) who had the closet "predicted earning" according to the prediction equation. The respondent whose earnings were still top-coded was then assigned the imputed or observed earnings of his/her nearest neighbor match. In this algorithm, a respondent's nearest neighbor match might not be the same respondent over time.

# 6 Special Cases

## 6.1 Respondents who Never Had Any Earnings

A respondent who consented but never had any earnings would have all zeroes in summary records but no entry at all in the detailed records. There are 694 such respondents. We only had one row of record for each of them in this data, with most variables set to '.m". This allows the researchers do distinguish them from respondents with zero earnings, or in respondent-years when earnings records were not available. See Section 7 for more details.

[4] We used inverse sine transformation, $\ln(\sqrt{Y^2+1}+Y)$, rather than $\ln Y$, to include respondents with zero earnings in the regression.

[5] To calculate the cumulative covered earnings, earnings above the taxable cap were excluded.

### 6.2 Inconsistency in Consent History across Detailed and Summary Records

For each respondent, the HRS received from SSA the earnings history from 1951 (summary records) or 1978 (detailed records) through the last year that the most recent consent was applicable. In the early years, the consent was only valid for that specific wave, and HRS had to obtain updated consent from the respondent in a later year in order to receive updated earnings records from SSA. Starting from 2006, the consent became "prospective" and remained valid until the late 2020s, eliminating the need for re-consent. Regardless the arrangement, there is a one-to-one relationship between the year of consent and the last year the earnings records were updated through.

In some cases, the "most recent consent" differed in summary and detailed records. In this dataset, we handled the inconsistency according to the following rules:

- We used the most recent consent date if the consent dates were different across sources. Most of these cases occurred after prospective consent went into effect, so the HRS would have received the same records anyway regardless when that consent was signed.
- For the very few cases where we had data beyond the years that a respondent's consent would have applied to, we used the last data point as "Last Year."

## 7 Variables in This Data Set

This data product is at the respondent-year level. Each respondent has only one observation in any given year. Except for the respondents who never had any earnings, the first record for a respondent is the first year the respondent had any positive earnings in SSA earnings records, and the last record for a respondent is the last year his/her most recent consent was applicable. Besides the household and person identifiers HHID and PN, this data product contains the following variables:

- Year: Calendar year of the earnings record. ".m" for 694 respondents who gave consent but did not show up in our records because they never had any earnings.
- CovPattern: Pattern of coverage, directly from summary records. Missing if the corresponding respondent-year record is from the detail record.

- TotQC: Total quarters of coverage, directly from summary records. Missing if the corresponding respondent-year record is from the detail record.
- NumRecords: Number of detailed earnings records available for the respondent in that year. It takes the value of non-zero positive integer except for the following cases:
  - 0, if the respondent had no earnings in a year when a detailed earnings record would otherwise have been available had there been any earnings.
  - .n, if earnings information for the respondent-year was obtained from years not covered by detailed earnings records.
  - .m, for the 694 respondents who never had any earnings.
- CovY: Amount of earnings in OASDI-covered employment. The number is non-negative except for the following cases:
  - missing, if there was no earnings record at all for that respondent-year.
  - .m, for the 694 respondents who never had any earnings.
- CovYuptoCap: Amount of earnings in OASDI-covered employment, up to the OASDI taxable cap. The number is non-negative, with the same exceptions as "CovY."
- CovYabvCap: Amount of earnings in OASDI-covered employment above the OASDI taxable cap. The number is non-negative, with the same exceptions as "CovY."
- NonCovY: Amount of earnings in employment not covered under OASDI. The number is non-negative except for the following cases:
  - missing, if there was no detailed earnings record for that respondent-year.
  - .n, if the record came from a year before detailed records were available.
  - .m, for the 694 respondents who never had any earnings.
- SelfEmpY: Amount of earnings in OASDI-covered self-employment. The number is non-negative except for the following cases:

  - missing, if there was no detailed earnings record for that respondent-year.
  - .n, if the record came from a year before detailed records were available.
  - .m, for the 694 respondents who never had any earnings.

- CovYTopCoded: A flag indicating whether covered earning was top-coded at the OASDI taxable maximum in the original data that HRS received from SSA.
  - 0, if the earnings were not top-coded.
  - 1, if the earnings were top-coded.
  - missing, if there were no earnings record from the respondent-year.
  - .m, for the 694 respondents who never had any earnings.

- SETopCoded: A flag indicating whether covered self-employment earning was top-coded at the OASDI taxable maximum in the original data that HRS received from SSA. Its values have the same definition as "CovYTopCoded".

- ImpFlag: A flag indicating whether the earnings numbers were imputed due to top-coding.
  - 0, if no imputation, except for scenarios specified as 1 or 2 in DERFlag below.
  - 1, if the respondent-year was from summary records and pattern of coverage was available.
  - 2, if the respondent-year was from summary records and pattern of coverage was not available.
  - 3, if the respondent-year was from detailed records and self-employment earning was top-coded.
  - missing, if there were no earnings record from the respondent-year.
  - .m, for the 694 respondents who never had any earnings.

- DERFlag: A flag indicating whether the earnings number from detailed record was imputed.

  - 0, if no imputation.
  - 1, if earning was above $250,000 and hence masked as ".x", ".y", or ".z".
  - 2, if earning from detailed record was above $250,000 but the earning from summary record in the same year was not top-coded. This likely is caused by mistakes in SSA's Master Earnings File, and the earning in this respondent-year file was replaced with the earning reported in summary record.
  - missing, if the respondent-year was not covered by detailed records.
  - .m, for the 694 respondents who never had any earnings.

Additionally, the following variables are at respondent- or year-level.

- FirstRec: First year when a positive earnings was reported.
- LastRec: The most recent year when a positive earnings was reported for the respondent.
- Consent: The most recent consent the HRS has for the respondent.
- LastYr: The last year the most recent consent applies to for the respondent.
- OASDICap: OASDI taxable cap of the year.
- HICap: Medicare taxable cap of the year, only available in 1991-1993 and missing otherwise.
- AWI: Social Security national average wage index of the year.

## References

**Feenberg, Daniel R., and James M. Poterba.** 1993. "Income Inequality and the Incomes of Very High-Income Taxpayers: Evidence from Tax Returns." *Tax Policy and the Economy*, 7: 145–177.

**Kopczuk, Wojciech, Emmanuel Saez, and Jae Song.** 2010. "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data since 1937." *Quarterly Journal of Economics*, 125(1): 91–128.

**Olsen, Anya, and Russell Hudson.** 2009. "Social Security Administration's Master Earnings File: Background Information." *Social Security Bulletin*, 69(3): 29–45.

**Piketty, Thomas, and Emmanuel Saez.** 2003. "Income Inequality in the United States, 1913-1998." *Quarterly Journal of Economics*, 118(1): 1–41.