Health and Retirement Study
Imputation of Lifetime Earnings

Data Description and Technical Documentation

Chichun Fang
Institute for Social Research
University of Michigan

Version 2022
April 2026

Contents

1 Overview

This data product, *Imputation of Lifetime Earnings*, provides lifetime earnings estimates of respondents in the Health and Retirement Study (HRS). It uses information from earnings records provided by the Social Security Administration (SSA), imputation, as well as projection, to estimate cumulative lifetime earnings through various ages. This documentation details the types of SSA earnings records available to the HRS, how the earnings records are utilized in this data product, and what imputation and projection methods are used when SSA earnings records are not available. Users who only need to know the definitions of variables included in this data product may proceed to Section 4 directly.

This restricted data product 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*. 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.

2 Earnings Records Available to the HRS

The HRS asks its respondents to provide consent to allow their SSA administrative records to be linked to their survey responses for research use. The consent form includes information such as name, birth date, and Social Security number, which SSA uses to match with its master earnings file (MEF) and master beneficiary records (MBR). Before 2006, we asked respondent consent each time they were interviewed, and the administrative records (if matched by the SSA) were updated through the most recent consent year. Since 2006, the consent has become "prospective." Each respondent would only have to consent once, and the consent remains effective throughout the duration covered by the agreement between HRS and SSA.

Not all respondents consented, and not all those consented can be matched by the SSA. For those who have matched SSA earnings records, we use earnings updated through 2022 (for respondents

who provided prospective consents) or the most recent year available (for respondents who consented before 2006 but did not provide prospective consents afterwards) to construct lifetime earnings estimates.

Two types of SSA administrative earnings records are available to the HRS: detail earnings records and summary earnings records. The following describes the time range that each type of record covers as well as information available in them.

2.1 Detail Earnings Records

Detail earnings records (DERs) are available for earnings years 1978 onward. DERs are at the respondent-year-employer level. In general, there are two types of DERs: DERs corresponding to employment earnings (that is, wages paid by employers) , and DERs corresponding to earnings from FICA-covered self-employment or tips.

DERs corresponding to employment earnings are available regardless whether the employment is covered by Social Security.[1] Each of such DERs contains three earning amounts: earnings subject to federal income tax (that is, the number reported in W-2 Box 1), earnings subject to Social Security tax (W-2 Box 3) and earnings subject to Medicare tax (W-2 Box 5). Earnings subject to federal income tax are not top-coded.[2] Earnings subject to Social Security and Medicare taxes are top-coded at their respective taxable maximums.[3] Each of the DERs corresponding to earnings from FICA-covered self-employment or tips contains two earnings amounts: earnings subject to Social Security tax and earnings subject to Medicare tax, both top-coded at their respective taxable maximums.

How we utilize the earnings information reported in DERs depends on the varying taxable maximums (and hence top-coding) since 1978, which is discussed in Appendix A.

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

[2] Per the agreement between HRS and SSA, earnings above $250,000 are "masked" in the DERs that the HRS provides as a restricted data product for confidentiality reasons. For high earners in DERs, approved HRS restricted data users can only see the range, but not the exact amount, of earnings. In this data product, however, we use the unmasked numbers to calculate lifetime earnings.

[3] For wages received in public sector jobs not covered by Social Security, earnings subject to Social Security will be zero. However, most employees in such jobs still pay into Medicare.

2.2 Summary Earnings Records

Summary earnings records (SERs) are available for earnings years 1951 onward. Each respondent only has one SER, which contains earnings in each year between 1951 and 2022 as well as the the pattern of coverage in each year. Only earnings in FICA-covered jobs are reported in SER, and earnings are top-coded at the FICA taxable maximum in each year. Such top-coding increases the need for imputation; additionally, SER contains no information about earnings outside FICA-covered employment. Hence, we only use SER information between 1951 and 1977, the period when DERs are not available. From 1978 onward we rely on DERs.

We use the pattern of coverage information to impute earnings top-coded in SER. Between 1951 and 1977, employers were required to report quarterly earnings information to the SSA, and pattern of coverage in a given year summarizes whether earnings were reported in each quarter of that year. In Appendix B, we explain in detail what the pattern of coverage is and how we use the information to impute top-coded earnings.

3 Data Structure Used to Calculate Lifetime Earnings

To calculate lifetime earnings, we first construct a panel of respondent-age/year earnings data. We use linked SSA earnings records, imputation, and projection to "fill in" this respondent-age/year panel.

We call a respondent in the "matched sample" if the HRS has at least some linked SSA earnings records for him/her. A respondent whose earnings records were never made available to the HRS is in the "unmatched sample." The discussion in the following subsections is organized by whether a respondent is in the matched or unmatched sample, as well as the age/year when the earnings records are available.

3.1 Determining Lifetime Earnings for the Matched Sample

If a respondent's earning records are available through age 70 or calendar year 2022 (whichever earlier), we use all the SSA earnings records for this respondent.

For respondents whose earnings records are available through 2022 but have not attained age 70 as of 2022, we project his/her earnings beyond year 2022 and through age 70 using the method detailed in Section 3.3.

For respondents who consented before when the prospective consent went into effect but never re-consented, the HRS has their earnings records through 1992, 1998, or 2004 (depending on when the most consent was for each respondent). How their earnings records are used depends on the respondent's age when the most recent year the records are available. If a respondent had already reached age 70 when the records ended, we use all the records through age 70. If If a respondent had not reached age 70 when the records ended, we impute the earnings through age 70 or year 2022 (whichever earlier) using a method similar to what is detailed in Section 3.2. If such a respondent still was not aged 70 as of 2022, we then utilize the projection method detailed in Section 3.3.

Hence, for the matched sample, the respondent-age/year panel stared in the year when the very first SSA earnings record was available and ended in age 70 or year 2022 (whichever earlier).

3.2 Determining Lifetime Earnings for the Unmatched Sample

We impute the earnings records for the unmatched sample. The imputations are done in two stages. First, we impute the cumulative lifetime earnings at the time of HRS entry. Second, conditional on the imputed cumulative earnings at HRS entry, we iteratively impute yearly earnings using information available in the HRS core survey.

The cumulative lifetime earnings at the time of HRS entry for the unmatched sample is imputed using the nearest neighbor matching. For the matched sample, the (inflation-adjusted) lifetime earnings at HRS entry can be calculated directly using the linked SSA earnings records.[4] For the unmatched sample, this variable is missing. Using the matched sample, we run prediction equations that regress the log of lifetime earnings[5] on the following variables: race (minority or not), ethnicity (Hispanic or not), level of education dummy variables (high school dropout [omitted category], high

[4]Regardless timing and type of consent, all respondents in the matched sample would always have linked earnings records through HRS entry. The type and timing of consent only affect whether earnings records in years *after* HRS entry are available.

[5]We use the inverse hyperbolic sine transformation, $ln(y + \sqrt{y^2 + 1})$, rather than $ln(y)$, for the logarithm of earnings and income variables, so the observations with zero or negative values are retained

school graduate, some college, college, and more than college), whether the respondent was born in the U.S., marital status at the baseline interview, labor force status at the baseline (work full-time [omitted category], work part-time, unemployed, partially retired, retired, disabled, and not in the labor force other than retirement), individual annual income at the baseline, household assets at the baseline, and a set of entry age dummy variables. The prediction equations are estimated separately by gender and by HRS cohort. Using the regression coefficients, we then estimate the predicted value of cumulative earnings at HRS entry for both the matched and the unmatched sample, conditional on the same sets of right hand side (RHS) variables. For each respondent in the unmatched sample, we find a respondent in the matched sample that has the most similar predicted cumulative earnings at HRS entry, and use this matched respondent's actual lifetime earnings at HRS entry as the unmatched respondent's imputed lifetime earnings at entry.

The yearly earnings are also imputed using a similar nearest matching method. All the time-varying variables on the right hand side of the regression model described in the previous paragraph are now at the corresponding wave/year of the imputation.[6] There are also a few more variables in the prediction equation: log of wage or earnings for that calendar year[7], log of cumulative earnings at HRS entry, a set of age dummy variables, and HRS cohort dummy variables. The left hand side variable of the prediction equation is total annual earnings according to the linked SSA earnings records. The regressions are estimated for each calendar year and by gender. Again, we use the coefficients to calculate predicted yearly earnings for both the matched and the unmatched sample; for each respondent in the unmatched sample, we find his/her "nearest neighbor" in the matched sample and use the nearest neighbor's actual yearly earnings as the unmatched respondent's imputed lifetime earnings.

Hence, for the unmatched sample, the respondent-age/year panel stared in the age/year of HRS entry and ended in age 70 or year 2022 (whichever earlier). The first record (at the year of HRS entry) contains the cumulative lifetime earnings through the year of the HRS entry, and the rest

[6]Note that the HRS is a biennial survey. Hence, except for wages/earnings, the values of other time-varying self-report variables for year t are also used for year $t+1$.

[7]The HRS asks both wages in the current job and earnings from the previous year. Hence, for each calendar since the HRS entry, we have information on either wages or earnings.

records contain the annual earnings at each corresponding age. For respondents in the unmatched sample who entered the HRS after age 70, his/her cumulative earnings at HRS entry is considered as his/her lifetime earnings through age 70 (see Section 4 for more details).

Finally, for the matched sample who did not provide prospective consents, we may not have enough linked SSA earnings records to calculate lifetime earnings at later ages. We impute their earnings in the years where records are missing as if they are in the unmatched sample in those years. Their actual (i.e. SSA-matched) earnings records, when available, are still retained, and imputations are only used in the years when there were no valid consents.

3.3 Projecting Future Earnings

We use the same projection algorithm (Mitchell, Olson and Steinmeier, 2000) as those in the data product *Cross-Wave Prospective Social Security Wealth Measures of Pre-Retirees* to project earnings for the respondents who have not attained age 70 as of 2022. Earnings in the five-year period preceding the wave year (t) are indexed to the $t-1$ levels using the national Average Wage Index (AWI). The indexed wages are then averaged, with years $t-1$ through $t-5$ single year earnings given weights 5, 4, 3, 2, and 1, respectively.

The projected real earnings for a future year t can be expressed as:

$$Y_t = \frac{1}{15}\Big(1+CPI_t\Big)\Big(5\cdot Y_{t-1} + 4\cdot Y_{t-2}\cdot\frac{AWI_{t-1}}{AWI_{t-2}} + 3\cdot Y_{t-3}\cdot\frac{AWI_{t-1}}{AWI_{t-3}} + 2\cdot Y_{t-4}\cdot\frac{AWI_{t-1}}{AWI_{t-4}} + 1\cdot Y_{t-5}\cdot\frac{AWI_{t-1}}{AWI_{t-5}}\Big)$$

where Y_t represents the earnings in year t and $\frac{AWI_{t-1}}{AWI_{t-i}}$ is the index factor that inflate earnings in year $t-i$ to year $t-1$ dollars. By definition, AWI is not available beyond 2022. AWI beyond 2022 is estimated using the assumed wage growth rate under the intermediate economic assumptions in the Trustees' Report. For each respondent under age 70, the projection equation is repeated iteratively through the year s/he attains age 70.

4 Variables in This Data Set

Cumulative lifetime earnings at various ages can be calculated using the respondent-age panel described in Section 3. Records in this data product are at the respondent level. "Lifetime earnings"

is defined as the sum of real (inflation-adjusted) earnings at each age. Yearly earnings may come from matched Social Security earnings records, be imputed, or be projected using the methods described above. All the earnings numbers are expressed in 2022 USD and are rounded to the nearest $1,000.

We are not able to impute lifetime earnings at HRS entry for respondents who do not have linked SSA earnings records *and* whose year of HRS entry is missing in the tracker file. These respondents are not included in this data product.

- HHID: Household identification number used in the HRS.
- PN: Person number used in the HRS.
- LE50: Lifetime earnings through age 50. Missing for HRS who are not matched and entered the study after age 50.
- LE55: Lifetime earnings through age 55. Missing for HRS who are not matched and entered the study after age 55.
- LE60: Lifetime earnings through age 60. Missing for HRS who are not matched and entered the study after age 60.
- LE65: Lifetime earnings through age 65. Missing for HRS who are not matched and entered the study after age 65.
- LE70: Lifetime earnings through age 70. Alternatively, for unmatched HRS respondents who entered the study after age 70, their imputed lifetime earnings at HRS entry.
- VERSION: Version number of this data release.

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.

Mitchell, Olivia S., Jan Olson, and Thomas L. Steinmeier. 2000. "Social Security Earnings and Projected Benefits." In *Forecasting Retirement Needs and Retirement Wealth.* , ed. Olivia S. Mitchell, P. Brett Hammond and Anna M. Rappaport, Chapter 13. University of Pennsylvania Press.

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.

A Construction of the Imputed Earnings Records Data

We describe how earnings are constructed/imputed in each period in a reverse chronological order, as more records were masked or top-coded in earlier years.

To understand the procedures detailed in this section, the user needs to know the type(s) of SSA earnings records available to the HRS. For most respondents, summary earnings records (SERs) are available between 1951 and 2022, and detail earnings records (DERs) are available from 1978 onward. Due to the nature of data, the level of complexity to construct earnings data goes up further back in time. In the following, we explain how the earnings information is constructed or imputed in a reverse chronological order, so the discussion starts from the least to the most complicated periods.

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 SERs but no DERs for them, and the records stopped in the year the consents were given. For these respondents, their earnings are treated using the same procedure as in "between 1951 and 1977" subsection below, regardless which year the earnings record(s) corresponded to.

A.1 Earnings from 1994 Onward

We only rely on DERs for earnings in this period. For both covered and non-covered employments, wage earnings in detail earnings records are not top-coded. For covered self-employment, earnings subject to Social Security tax are top-coded at OASDI taxable maximum, and earnings subject to Medicare tax are top-coded at Medicare taxable maximum. However, Medicare taxable maximum has been suspended since 1994, so self-employment earnings from 1994 onward could be treated as not being top-coded as well.

Earnings in a given year are operationally defined as the maximum of (a) and (b), where:

- (a) is the sum of two components: (a1) earnings and tips subject to federal income tax, and (a2) contributions to qualified tax-deferred plans in the form of payroll deduction;
- (b) is earnings and tips subject to Medicare tax.

We have components (a) and (b) for employment earnings (covered or not); in most cases (a) and (b) should be the same given that employees in most non-covered jobs are also required to pay Medicare tax. We only have (b) for covered self-employment earnings.

A.2 Earnings between 1978 and 1993

We also only rely on DERs for earnings in this period. This period was the same as 1994 onward with one exception: the Medicare taxable maximum was in effect, so self-employment earnings subject to Medicare tax reported in DERs was effectively top-coded. Between 1991 and 1993, the Medicare taxable maximum was higher than the Social Security taxable maximum, which means Medicare-taxable earnings were less likely to be top-coded than Social Security taxable earnings. Between 1978 and 1990, Medicare taxable maximum and Social Security taxable maximum were the same. We use the nearest neighbor matching method in Appendix D.2 to impute the earnings of self-employed respondents that were top-coded.

The Medicare taxable maximum during this period did not affect earnings from covered or non-covered employment reported in DERs. While earnings subject to Medicare tax were still top-coded, earnings subject to federal income tax were not. For employment earnings between 1978 and 1993, we still use component (a) discussed in the previous subsection.

A.3 Earnings between 1951 and 1977

We only have earnings reported in SERs for the period between 1951 and 1977. As previously mentioned, for 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.

SERs differ from DERs among several aspects: (1) only earnings records from covered employment or covered self-employment are available; (2) in each year for each respondent, all covered earnings are lumped together and reported as one single number; and (3) the reported covered earnings are top-coded at the Social Security taxable maximum. They lead to two important restrictions in this data product: first, non-covered earnings are not available during this period, and we do not attempt to impute non-covered earnings; second, we do not attempt to separate earnings

from covered employment and covered self-employment.

Except for the respondents who only consented in 1992 or 1993 but never re-consented later, a "pattern of coverage" variable is available in the SERs.[8] With some assumptions regarding earnings patterns, this variable can be used to infer which quarter in a calendar year the taxable maximum was hit. The assumption we make are explained in Appendix B.

For the top-coded records in respondent-year that we know when the taxable maximum has been hit (through the "pattern of coverage" variable), we use Pareto or Log-Normal interpolations to impute the earnings. The parametric assumptions and procedures of imputation are detailed in Appendices C.1, C.2, and D.1. That leaves us the earnings of respondents whose pattern of coverage information is not available or who were self-employed (self-employed earnings were reported to SSA annually rather than quarterly). Covered earnings of these respondents are imputed using nearest neighbor matching as detailed in Appendix D.2.

B Pattern of Coverage in Summary Earnings Records

B.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

[8] For respondents who only consented in 1992/1993, HRS does not have their pattern of coverage information due to our agreement with SSA.

indicates 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 after the taxable maximum was reached.

Pattern of coverage information was available for most HRS respondents between 1951 and 1977 with two exceptions. First, for 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, their consent forms included neither detailed earnings records nor the pattern of coverage variable in summary records. In such cases, the pattern of coverage was "9999", indicating "not applicable". We explain how alternative imputation strategies are developed for this group in Appendix D.2.

Secondly, 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 employer's 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 employer's 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 do not affect our imputation. We consider the pattern "0000" as pattern of coverage not available and treat them the same as "9999" in Appendix D.2.

B.2 Assumption about the Earnings Pattern

The OASDI taxable maximum is Y_t^* in year t. Y_{it} is the earnings of individual i in year t. In years when earnings reported in SSA data were top-coded, 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 assume that, for these respondents, annual earnings distributed uniformly across four calendar quarters.

This assumption allows us to determine the range of possible annual earnings relative to Y^* (the subscript t is omitted for simplicity) 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^*. The same assumption also implies that 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.

C Parametric Assumptions about the Earnings Distribution

Consistent with the earnings distribution literature, we assume that the earnings distribution has two "segments." In any given year, the earnings distribution is log-normal up until a certain threshold and becomes Pareto beyond that threshold. We define such threshold as 200% of Y^*. In other words, in any given year, those who hit Y^* in Q1 or Q2 have earnings following a Pareto distribution, and the rest of sample have earnings following a log-normal distribution.[9] Such definition of the threshold is based on the actual earnings patterns observed in the data. Between 1951 and

[9]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 are not able to do so as we do not have the same quarterly earnings data that they had.

1977, about 1% of respondents hit Y^* in Q1 or Q2. There are 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.

C.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 α:

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

$K > 0$ is the minimum level of earnings to which the Pareto distribution applies, and α 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 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 allow $\hat{\alpha}$ 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 $\hat{K}$ are generally higher than those reported in Feenberg and Poterba (1993) but comparable to those reported in Picketty and Saez (2003).[10]

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

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

[10] Both Feenberg and Poterba (1993) and Picketty and Saez (2003) used data that were representative of all U.S. taxpayers. However, our data are not supposed to be representative of all U.S. taxpayers between 1951 and 1977 (or at any point of time for that matter) because by construction HRS respondents are not representative of all U.S. taxpayers. Hence, we do the comparison only to assess how the distributions compare with other results rather than to check the validity of our estimates.

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.

C.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 assume the earnings to be log-Normally distributed.

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, $Y_{Q3} = 1.33 \cdot Y^*$ and $Y_{Q4} = Y^*$ are the minimum earnings required in order to hit the taxable cap by Q3 and Q4, respectively. By the 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 μ and σ 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 allow $\hat{\mu}$ and $\hat{\sigma}$ to vary over time, and they are calculated with Y_{Q3} and Y_{Q4} being replaced by year-specific $1.33 \cdot Y_t^*$ and Y_t^*, respectively. With $\hat{\mu}$ and $\hat{\sigma}$, earnings at any given percentile within the range can be calculated.

D Imputation Strategy

D.1 Pareto and Log-Normal Interpolations

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

In years when only SERs were available, the assumption in Section B.2 allows us to "bracket" where a respondent was in the earnings distribution. For example, say there are 960 respon-

dents/records that are not top-coded, 20 respondents hitting the earnings maximum in Q4 (and hence between 100% and 133% of Y^*), 10 hitting the maximum in Q3 (between 133% and 200% of Y^*), 7 hitting the maximum in Q2 (between 200% and 400% of Y^*), and 3 hitting the maximum in Q1 (above 400% of Y^*). We can infer that those who hit the maximum in Q4 must be between 96^{th} and 98^{th} percentiles, those who hit the maximum in Q3 must be between 98^{th} and 99^{th} percentiles, those who hit the maximum in Q2 must be between 99^{th} and 99.7^{th} percentiles, etc.

This is the first step of our imputation, which brackets 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 are 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 are top-coded in year 1 and 10% are top-coded in year 2. Using the information in year 1, we can calculate where everybody is in the ranking using the non-top-coded earnings up to the 99^{th} percentile. We can also know who are among the top 1%, although we do not know exactly where they are along the distribution. If we assume the relative ranking stays the same in both years, these two pieces of information allow us to know the ranking of each respondent up to the 99^{th} percentile as well as who are among the top 1%. Consequently, although 10% respondents have earnings top-coded in year 2, we could still infer where they are along the distribution and impute the earnings accordingly.

In this data product, we relax the assumption that relative ranking stayed the same over time. We instead assume that rankings are "smooth" and predict individual ranking with a moving average. The ranking is done by HRS cohort, to account for the correlation between earnings rank and age. There are also enough granularities in the data to further allocate respondents into the four categories of top-coding: those who hit the maximum in Q1, Q2, Q3, or Q4 between 1951 and

1977. This is the second step of our imputation.

Finally, we rank the predicted ranking from Step 2 within the "bracket" of ranking obtained in Step 1. That is, for example, 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 is 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 preserve the "bracket information" (the range of earnings distribution where the respondent belonged to) and supplement it with earnings ranking from adjacent years.

D.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 estimate a prediction equation with all the respondents whose (a) earnings are below the cap, (b) earnings are top-coded at the cap in SERs but imputed using the method in Appendix D.1, or (c) earnings are at or above the cap but directly available in the DERs. The prediction equation is estimated separately by year and HRS cohort.

The left-hand side of the prediction equation was the logged earnings in year t.[11] The right-hand side includes the cumulative covered earnings between 1951 and year t (price-adjusted using the Social Security average wage index)[12], 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 is estimated separately by year and

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

[12] For the purpose of this prediction equation, earnings above the taxable cap are excluded.

HRS cohort. The coefficients from the prediction equation are used to calculate a "predicted earning" for everybody, pattern of coverage information available or not.

Each respondent whose earning is still top-coded at this stage is matched to a respondent (a) whose earnings are above the taxable cap and has been imputed previously in Appendix D.1 or directly observed, and (b) who has the closet "predicted earning" according to the prediction equation. The respondent whose earnings are still top-coded is 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.