Adding Historical Earnings to the Survey of Income and Program Participation (SIPP)

Karen E. Smith, Fritz Scheuren, and Jillian Berk,
Urban Institute, 2100 M Street NW, Washington DC, 20037

KEY WORDS: Imputation, Statistical Match, Historic Earnings.

1. Introduction

Social Security under current law is insolvent and the system will require reforms to protect family well-being in the future. President Bush has suggested reform that would include private accounts, and dozens of Senators and Congresspeople have advanced plans to respond to the difficulties of this popular program. In the past few years, concerned citizens, advocacy organizations, and research institutions have joined in the ongoing debate about the future of the federal retirement income program. These concerns have fueled the need to have public access to data sets and computer models that will allow the modeling of the distributional effects of alternate Social Security policy regimes.

The Social Security Administration (SSA) is interested in giving researchers better access to its administrative data, however, much of this data is covered by confidentiality agreements and cannot be made publicly available. We have created a public use version of the Social Security Administration's historic earnings records matched to the 1990 to 1993 SIPP panels for the Urban Institute DYNASIM microsimulation model. This paper describes the statistical matching method we used to generate this public use file. It then compares the imputed file with actual administrative data to show that the imputed file has a high degree of similarity to the actual data by cohort, gender, race, education, and marital status.

We created historic earnings by statistically matching earnings from the Panel Survey of Income Dynamics (PSID) and a public use version of the 1973 Current Population Survey (CPS) matched to the Social Security Summary Earnings Records (SER). After the statistical match, we have a fully public use file of over 180,000 SIPP individuals with historic earnings from 1951 (the first year the Social Security Administration collected annual earnings in the summary earnings record) to 1993.

Our statistical matching procedure is based on minimizing a distance function weighted by the partial r-square generated from a stepwise regression with lifetime earnings as the dependent variable. We compare this methodology with a predicted mean match, a commonly used technique for statistical matching (Haider et al (2000), O'Hare (1997)), and discuss why we believe the partial r-square method is superior. Finally, we compare lifetime earnings from our imputed file with actual lifetime earnings observed on the administrative data. We show that the imputed file does a very good job of capturing lifetime earnings by cohort, gender, education, marital status, and race.

We believe that the very good results from our statistically matched file validate the earning histories used in the DYNASIM file. At the conclusion of the two statistical matches, we have a file that achieves a high degree of verisimilitude or "look-likeness," yet differs enough from the administrative records that it can be released as a public use file without compromising the confidentiality of the data. Thus, the weaknesses of this file are actually its greatest strength. We plan to release the statistically matched file so that researchers without access to the administrative data will have a starting point for modeling the distributional effects of proposed Social Security reforms.

In section 2 of this report, we briefly describe the data files we used in the statistical matches. In section 3, we describe the statistics we use to measure lifetime earnings. This measure is the basis for defining the distance function weights and for evaluating the results. In section 4, we describe our weighted r-square methodology and contrast it with the predicted mean match. In section 5, we show our results, and in section 6, we present our conclusions.

2. Data

SIPP Data: The SIPP is a continuous series of national panels of noninstitutionalized households beginning in 1984, with sample size ranging from approximately 14,000 to 36,700 interviewed households. The survey collects detailed information on incomes by source, labor force participation, and demographic characteristics. It also collects periodic information on assets, pensions, marital and fertility history, health and functional limitations, among others.

1973 CPS/SER: The CPS is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. It collects information on hours of work, earnings,

other income sources, and a variety of demographic characteristics including age, sex, race, marital status, and educational attainment. The Census Bureau matched the 1973 March CPS with the Social Security Summary Earnings Record (SER) using the individual's social security number. This is the last publicly available cross-sectional household file with administrative earnings record the Census Bureau has released. The SER includes only earnings from Social Security covered employment and these earnings are capped at the Social Security taxable maximum.

PSID Data: The PSID is a longitudinal nationally representative sample of 6,000 households that began in 1968. All members of the original household are interviewed annually along with their spouses. The survey is ongoing with the most recent available interview being done in 1994. This gives 26 years of information for individuals in the original population sample. The survey collects substantial detail on income sources and amounts, employment, family composition, and residential location. The PSID includes uncapped earnings of heads and wives in covered and uncovered employment.

3. Measurement of Lifetime Earnings

We are interested in generating historic earnings that have a high degree of verisimilitude or "look-likeness" to actual earning histories. We want the imputed data to reflect actual labor force participation, level of earnings, and growth rates in earnings by age and education level. We are particularly interested in capturing the dramatic increases in female labor force participation and earnings that have occurred over the last several decades and the changes in retirement and disability patterns of men (Smith (2000), Fullerton (1999), Hayghe (1990), Wetzel (1990)).

We measure lifetime earnings using the average wage-indexed annual earnings from age 25 to age 65 (AIE). This is calculated as follows:

$$AIE = \sum_{y=b+25}^{b+65} e_y / a_y,\ firstyear <= y <= lastyear \qquad (1)$$

where b is birth year, e is annual earnings, a is the economy-wide average earnings, and y is year. We use this measure to capture very dynamic lifetime profiles that make up individual earning histories using a single statistic. In all cases, the range of years depends on the years available in the data source. On the PSID, years range from 1968 to 1993. For the CPS/SER, years range from 1951 to 1972. On the final matched file, years range from 1951 to 1993. Because of differences in the year ranges in each dataset, AIEs are not comparable across datasets. The AIE for each cohort within each dataset, however, gives a comparable measure of lifetime earnings.

4. Methodology

We do the statistical matching in two discrete segments: first, from the PSID to the base SIPP file. This gives us historic earnings from 1968, the first year of PSID data, to 1993. We then match this file to the CPS/SER to impute earnings from 1951 to 1967. The PSID match is based on observed earnings and demographic information from the SIPP interview and PSID characteristics in 1993 including age, race, family size, household wealth, annual earnings, pension income, and social security benefits. For the CPS/SER match, we match based on education, race, annual earnings from 1968 to 1972, and average earnings over that five-year period.

Our statistical matching procedure is based on minimizing a distance function. We limited the pool of potential donors to those individuals of the same gender born within 1 year of the desired birth year. Within the set of potential donors, we selected the "best" individual, where "best" is defined to be the individual with the smallest distance measured by a distance function. The distance function, in general, is defined as follows:

$$D_d = \sum_{j=1}^{n} w_j * [(X_{dj} - X_{rj}) / s_j]^2 \qquad (2)$$

where j is the number of measured attributes in the distance function, w is a weight factor, X is a characteristic measure, σ is the standard deviation of the jth X variable in the dataset, d denotes the characteristic of the donor, and r denotes the characteristic of the recipient.

The weight factor, w_j, allows the analyst to decide which attributes are more important to match on. In our applications, we used the partial r-square estimated from a regression with total earnings as the dependent variable. We calculated the distance, D, for each donor record, and selected the donor record with the smallest value.

We prefer the weighted distance function method to the predicted mean match method. In the predicted mean match, one calculates the predicted value of a desired variable. In our case, this would be total earnings. The match is then done by finding the donor record that minimizes the difference in predicted and actual value. This is equivalent to

equation 2 with only one X variable: predicted value for the recipient record, and actual value for the donor record. Here the weight would be one and the standard deviation unnecessary, because the X variables are on the same scale.

In the predicted mean match all variables in the prediction equation get equal weight in the match. If the donor sample is small and some of the variables in the prediction equation are poor predictors (have large standard errors), this lack of weighting can cause the predicted value to be extreme along some dimension that has little influence on the characteristic to be imputed.

After the statistical match is complete, we still need to adjust our imputed earnings to account for the limitations of the datasets that were used in the match. Our first problem is that the earnings vector constructed on the PSID is disproportionately censored at young ages. The PSID only ask heads and wives about their labor income. Individuals have a nonzero earnings value only for years in which they were a head or a wife. Since this does not usually occur until about age 25, the matched file has very low labor force participation for 15- to 25-year-olds imputed from the PSID. This is not a problem for older cohorts whose early life earnings are imputed from the non-age censored CPS/SER. For individuals born between 1943 and 1963, we match age 25 to age 30 wage-adjusted earnings from the SIPP/PSID file to age 25 to age 30 wage-adjusted earnings from the 1942 cohort of the CPS/SER. We effectively impute earnings from age 15 to age 30 from the CPS/SER to all cohorts born between 1943 and 1963.[1]

The final step is to impute Social Security covered employment and adjust for immigration. We assign Social Security coverage based on employment class observed on the SIPP. We assign all earnings for federal government employees covered by Civil Service Retirement System to uncovered employment. For state and local government employees, we randomly assign noncovered employment based on the state-specific rates of noncoverage.[2] Finally, for prior employment spells (earnings(t)=0 and earnings(t-1)>0) we randomly assign individuals to noncovered employment based on year-specific noncoverage rates.[3] Lastly, we zero out all earnings of new immigrants in years before they entered the United States (based on dates reported in the migration topical module on the SIPP).

5. Results

At the completion of these two statistical matches, we have a public use SIPP file with imputed lifetime earnings. To test the accuracy of the match, we compare the imputed public use file (we will call it DYNASIPP) with actual SIPP/SER matched data. The SIPP/SER file has the historical earnings records of all individuals who survive to the SIPP interview. They are the exact same sample of people that are in the imputed DYNASIPP file.

Comparisons of the distribution of AIE by cohort and gender show that the statistical match has done an excellent job capturing lifetime earnings along a number of dimensions (see Figure 1). Average AIEs compare quite well. The average imputed AIEs for men are about 6% too high for cohorts born after 1915. We may need to calibrate the assignment of uncovered work a bit to adjust these historic averages. The average imputed AIEs for women are very close for all cohorts.

Median AIEs are also very similar. Median AIEs are higher than mean AIEs for men, but lower for women and both track the actual data closely. For the 26th percentile, the imputed values for men are about 20% too high and almost identical for women. We may be understating the impact of uncovered earnings in the imputed data for men. When we impute Social Security covered employment, we generally assign all earnings of uncovered workers to zeros. In fact the pattern may be for more mixed spells of covered and uncovered employment. If we assigned more people part-career uncovered employment, this would lower the AIEs of more workers and thus lower the median AIE on the imputed data. At the 76th percentile of AIE, where uncovered employment is not an issue, the imputed AIEs are very close to the actual AIEs for both men and women. The imputed data also captures the variance in AIEs as shown by the standard deviation by cohort.

The imputed data also does a good job at capturing lifetime labor force participation as measured by the number of years with Social Security covered earnings greater than zero from age 25 to 65. Since most men in this age range work, the number of years worked largely measures the number of years each cohort attained age 25 and 65 between 1951 and 1993. The number is low for early cohorts, because they were at younger ages before 1951. The number is low for later cohorts, because they have not yet attained older ages by

[1] We zero out the weights in the distance function if the individual's earnings are missing due to not being a head or wife. The match uses only years with valid earnings, race, and education.
[2] Committee on Ways and Means (1996), Table 1-5 page 10-11.
[3] Committee on Ways and Means (1996), Table 1-2 page 8.

1993. For women, whose labor force participation rates are lower than men's, the imputed labor force participation compares exceedingly well to the actual participation. With the rising female labor force participation of later cohorts, the gap in men's and women's years worked is declining.

AIEs on the imputed data compare well with the actual data by marital status, race, and education as well (Figure 2). Naturally, as we disaggregate the data along additional dimensions, the data becomes much noisier, though the general closeness of the trends is still apparent. Married men have higher lifetime earnings for all cohorts compared to unmarried men. College graduate AIEs are higher than high school graduate AIEs and high school dropout AIEs. The imputed AIEs of college educated men born after 1955 are lower on the imputed file. Early career earnings for these records come from the CPS/SER where returns to education were not as high as today. Comparisons of male AIEs by race and cohort show that white men have higher AIEs than non-white men, but the difference is **declining** for later cohorts. Again, the imputed AIEs are very similar to the actual AIEs by race.

Single women have higher lifetime earnings for all cohorts compared to married women. This is the opposite relationship as for men. Married women have historically worked fewer hours compared to single women, though this trend is reversing over time. Comparisons of female AIEs by education and cohort show that college graduate AIEs are higher than high school graduate AIEs and high school dropout AIEs. Comparisons of female AIEs by race and cohort show that white women have higher AIEs than non-white women but the difference is **increasing** for later cohorts. This is the opposite trend compared to men. Again, the imputed AIEs are very similar to the actual AIEs by race.

6. Conclusions

Our statistical matching procedure based on minimizing a distance function weighted by the partial r-square does a very good job of imputing realistic-looking historical earnings from both the PSID and the 1973 CPS/SER to the SIPP. The technique allows for accurate imputations, even when the base distributions on the donor and recipient files are different, as is the case with covered earnings on the CPS/SER file and higher than average lifetime earnings on the PSID due to attrition bias.

Both lifetime earnings and number of years worked on the imputed file are very similar to those on the actual data. Comparisons by race, education, and marital status are also very similar. We believe the imputed data to be of very high quality along a number of critical dimensions making it a good starting data set for the DYNASIM model. We also believe the public use data file version available at www.urban.org to be of value to other Social Security modelers and those engaged in retirement research.

References:

Committee on Ways and Means. 1996. "1996 Green Book." U.S. House of Representatives. U.S. Government Printing Office, Washington D.C.

Fullerton, Howard N. Jr. 1999. "Labor Force Participation: 75 Years of Change, 1950-98 and 1998-2025" *Monthly Labor Review*, December 1999.

Haider, Steven, Michael Hurd, Elaine Reardon, and Stephanie Williamson. 2000. "Patterns of Dissaving in Retirement," AARP Public Policy Institute.

Hayghe, Howard V. 1990. "Family Members in the Work Force" *Monthly Labor Review*. March 1990: pp. 14-19.

O'Hare, John, 1997. "Impute or Match? Strategies for Microsimulation Modeling," A paper presented at Microsimulation in Government Policy and Forecasting International Conference on Combinatorics, Information Theory and Statistics. 1997.

Smith, Karen E. 2001. "The Status of the Retired Population, Now and in the Future." Presented at the Social Security and the American Family Conference, The Urban Institute, Washington DC.

Wetzel, James R. 1990. "American Families: 75 Years of Change" *Monthly Labor Review*. March, pp. 4-13.

Karen E. Smith is a Senior Research Associate, Fritz Scheuren is a Senior Fellow, and Jillian Berk is a Research Assistant at the Urban Institute.

The research in this paper was funded partly through the Mellon Foundation and the Center for Retirement Research at Boston College through a grant from the Social Security Administration. The views expressed in this paper are those of the authors and do not necessarily reflect those of the funders, the Urban Institute, its sponsors, or its trustees.

Figure 1

Actual and Imputed Lifetime Earnings (Ages 25-65) by Gender and Cohort

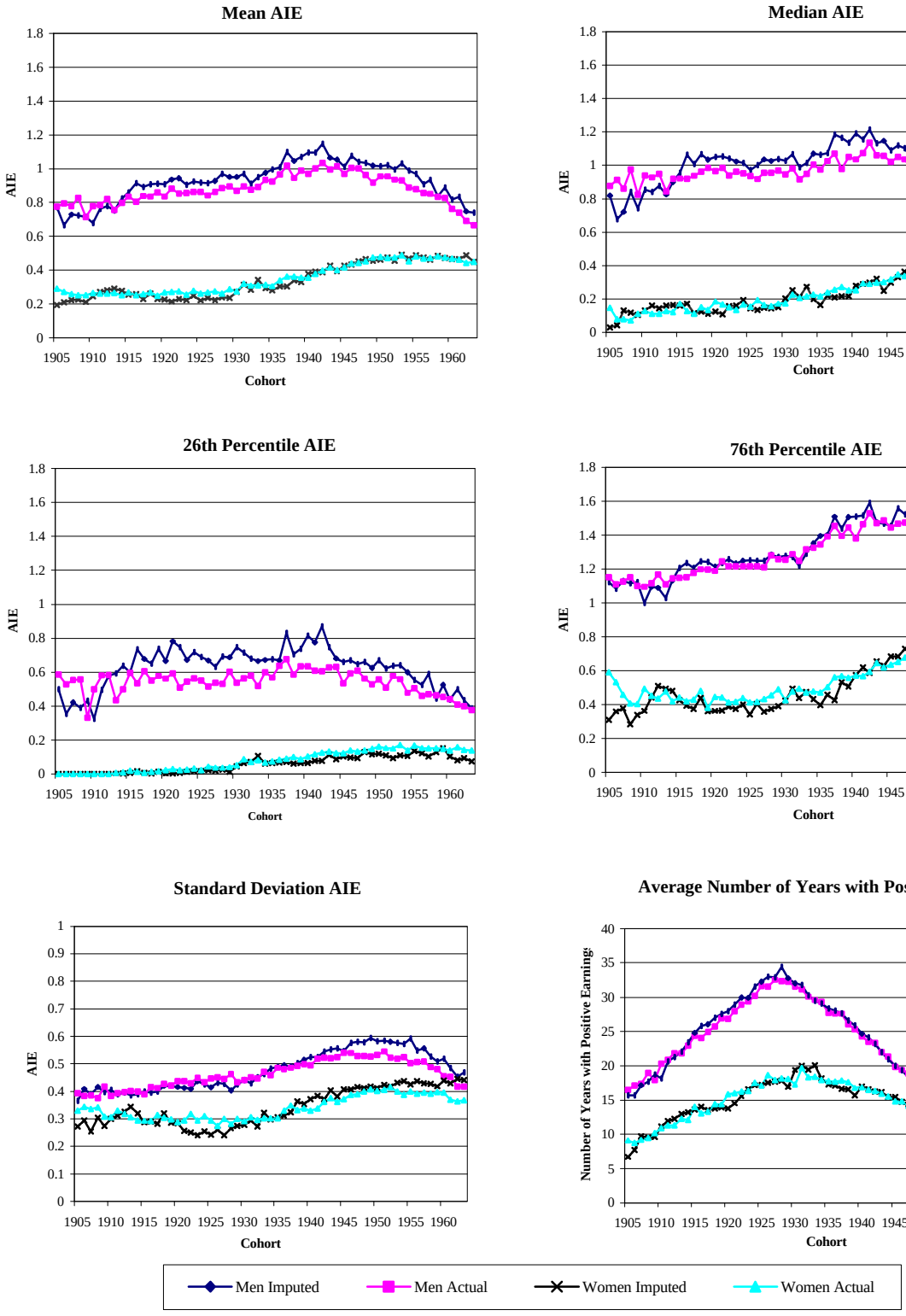

Figure 2

Actual and Imputed Median Lifetime Earnings (Age 25-65) by Gender, Race, Education, and Cohort