Nonrandom Selection in the HRS Social Security Earnings Sample

Steven Haider
RAND
1700 Main Street
Santa Monica, CA 90407
sjhaider@rand.org

Gary Solon
Department of Economics
University of Michigan
Ann Arbor, MI 48109
gsolon@umich.edu

February 2000

The authors gratefully acknowledge grant support from the National Institute on Aging (2P01-AG10179) and computing facilities provided by the University of Michigan's Population Studies Center and the Michigan Exploratory Center on the Demography of Aging.

Nonrandom Selection in the HRS Social Security Earnings Sample

I. Introduction

The Health and Retirement Study (HRS), administered by the Institute for Social Research (ISR) at the University of Michigan, is a longitudinal survey of the population of U.S. households with at least one adult between the ages of 51 and 61 in 1992 (individuals born between 1931 and 1941).[1] Wave 1 of the survey was carried out in 1992 with subsequent waves to be conducted every two years. In Wave 1, the HRS collected data on 12,652 respondents from 7,702 households.

In accordance with an agreement with the Social Security Administration, HRS respondents were asked to grant ISR permission to obtain the respondents' earnings histories as reported to the Social Security Administration. Such data are extremely valuable because they provide unusually accurate administrative earnings histories over an unusually long period, 1951-1991, and these data can be used in conjunction with the wealth of survey information collected in the HRS itself. Because of the highly confidential nature of the data, the earnings histories are not part of the HRS public release data sets; rather, the Social Security earnings data are provided only through special permission from the HRS.[2]

[1] See Juster and Suzman (1995) for a detailed description of the HRS.

[2] For additional information concerning the availability of these data, see the HRS website http://www.umich.edu/~hrswww/.

Although most respondents agreed to provide ISR access to their Social Security earnings records, some respondents refused.[3] As a result, the earnings histories are available for 9,472 respondents, 75 percent of the overall HRS sample of 12,652.[4] Thus, even though the HRS sample is intended to be nationally representative,[5] nonrandomness in the permission process could cause the sample of earnings histories to be unrepresentative. Fortunately, because the respondents that refused access to their earnings histories did respond to the HRS questionnaire, it is possible to explore whether the provision of permission is systematically related to the variables measured in the HRS. This note uses that information to explore the degree to which refusals to grant access to the Social Security data may have damaged the representativeness of the sample of earnings histories.

II. Tabular Evidence

A simple way to look for nonrandomness in the permission process is to examine whether the percentage with Social Security earnings histories varies with observable characteristics such as gender, race, education, wealth, and health. For example, with a "response rate" of 75 percent for the full sample, if the propensity to grant permission is independent of gender, then the separate response rates for women and men also should be close to 75 percent. Table 1 displays the percentages with Social Security earnings

[3] A few earnings histories were lost for other reasons, such as problems in matching the Social Security and HRS information. For simplicity, we fold these few cases in with the individuals whose earnings information is missing because they declined to permit access to their Social Security earnings records.

[4] Sixty-seven additional earnings histories were obtained for new sample members interviewed in Waves 2 and 3. These observations are omitted from our analysis.

[5] The HRS oversamples Hispanics, blacks, and Florida residents. Sampling weights are provided to enable analysts to adjust for the unequal probabilities of selection into the sample.

histories for a variety of subsamples. These sample response rates are accompanied by standard error estimates calculated as the square root of p(1-p)/N where p is the sample response rate and N is the number of sample observations in the category.

The response rates for many of the subsamples deviate only slightly from 75 percent.[6] In many instances, the deviations are so slight that the contrasts are statistically insignificant despite the large sample size. For example, the contrast between the 73.3 percent response rate for those that did not finish high school and the 74.0 percent rate for those that finished college is not significantly different from zero at any conventional significance level. Even when the contrasts are statistically significant, they often are quite modest. For example, the contrast by gender is statistically significant, but fairly small in magnitude – 73.8 percent for men versus 75.8 percent for women.

In a few instances, though, the contrasts are somewhat larger. For example, the response rate for those reporting that they never have worked is 67.7 percent as compared to 75.2 percent for those reporting that they have worked. It is not so surprising that those who never worked are more likely to decline permission to seek their Social Security records, as they have no covered earnings and may not even have a Social Security number. A similar point pertains to the six-percentage-point contrast between those born in the United States and those born elsewhere. Finally, there is a modest, but quite noticeable racial pattern, with blacks, Hispanics, and Asians all showing response rates below 70 percent.

[6] This finding accords with other researchers' preliminary analyses of an earlier version of the data set (Venti and Wise, 1998; Gustman and Steinmeier, 1999). We performed our analysis independently of Olson (1999), who recently reported results very similar to ours for the same data set. The main reason for the minor discrepancies between our results and Olson's is that we have not followed her practice of recoding missing values for a variable as belonging to the modal category.

Although some of the contrasts in Table 1 are statistically significant, it is reassuring that they are not terribly large. Furthermore, some of the largest contrasts may essentially be duplicates of each other, insofar as some of the characteristics examined are correlated with each other. Therefore, in the next section, we analyze *partial* relationships between the response rate and each characteristic with other characteristics held constant.

III. Logit Analysis

To examine partial relationships, we perform maximum likelihood estimation of a logit model for the probability π that the Social Security earnings history is available. The list of explanatory variables, their estimated logit coefficients, and the associated standard error estimates are shown in Table 2. The estimated logit coefficients can easily be translated into estimated effects on the response probability π by recognizing that the partial derivative of π with respect to a unit change in an explanatory variable is that variable's logit coefficient times $\pi(1-\pi)$. Thus, evaluated at a typical π of about 0.75, the estimated change in π with respect to a unit change in an explanatory variable can be obtained by multiplying that variable's estimated logit coefficient by 0.19.

At any conventional significance level, the chi-square statistic of 164.6 (with 17 degrees of freedom) easily rejects the null hypothesis that all the explanatory variables have zero coefficients. As in Table 1, however, most of the individual contrasts are small, and many are statistically insignificant. For example, the contrast between a respondent that completed college and one with less than a high school education is insignificant at the 0.10 level, and multiplying the coefficient estimate -0.120 by 0.19

indicates that the implied contrast is only about two percentage points. The estimated contrast between men and women is significant, and it amounts to about three percentage points. The estimated response probability is significantly increasing in household earnings and decreasing (at an absolutely decreasing rate) in household wealth, but the magnitudes of the estimated effects are small. Also, respondents that complain of poor health are slightly less likely to grant access to their Social Security records.

One of the largest contrasts in Table 1, between those born in the United States and those born elsewhere, mostly vanishes in Table 2 once other characteristics are controlled for. Some of the other contrasts, however, remain fairly substantial. The estimated contrast between those that reported never working and those that reported they did work remains statistically significant and is still about six percentage points. And even with other characteristics controlled for, nonwhites are estimated to be about eight percentage points less likely than whites to permit access to their Social Security earnings histories.

IV. Conclusion

About 75 percent of the HRS respondents have permitted access to their Social Security earnings histories. By and large, the availability of the Social Security information varies only weakly with the respondents' observable characteristics. Not too surprisingly, respondents that report they never worked are noticeably less inclined to offer access to their possibly nonexistent Social Security records. We also estimate that nonwhites are about eight percentage points less likely than whites to permit access.

Nevertheless, even these effects are not huge. As far as can be told from observable data, the HRS Social Security earnings sample seems to be reasonably representative.

Table 1
Percentage with Social Security Earnings Histories by Category

Category	Sample Size	Percentage with Social Security Earnings Histories	Standard Error
Total	12,652	74.9	0.4
Male	5,867	73.8	0.6
Female	6,785	75.8	0.5
Born by 1936	6,302	75.2	0.5
Born after 1936*	6,346	74.6	0.5
Born in U.S.	11,357	75.5	0.4
Born elsewhere*	1,294	69.7	1.3
White	9,112	77.1	0.4
Black	2,064	69.8	1.0
Hispanic	1,173	67.9	1.4
Asian	154	67.5	3.8
Native American	115	71.3	4.2
Other	6	83.3	15.2
Race not available	28	64.3	9.1
Did not complete high school	3,351	73.3	0.8
High school degree or equivalent	4,686	77.1	0.6
Some college	2,442	73.9	0.9
Completed college	2,090	74.0	1.0
Education not available	83	61.5	5.3
Retired	1,580	75.2	1.1
Not retired	11,072	74.9	0.4
1991 household earnings below $40,000	6,100	73.8	0.6
Earnings at least $40,000**	6,457	76.2	0.5
Never worked	496	67.7	2.1
Ever worked	12,156	75.2	0.4

(continued)

Table 1 (continued)
Percentage with Social Security Earnings Histories by Category

Category	Sample Size	Percentage with Social Security Earnings Histories	Standard Error
Reported government employment	875	75.1	1.5
Did not report government employment	11,777	74.9	0.4
Homeowner	9,849	75.1	0.4
Not homeowner**	2,708	74.8	0.8
Household net worth below $100,000	6,353	75.9	0.5
Net worth at least $100,000**	6,204	74.2	0.6
Reported excellent, very good, or good physical health	9,832	75.5	0.4
Reported fair or poor physical health	2,820	72.3	0.8
Reported job-limiting disability	2,717	75.1	0.8
Did not report job-limiting disability	9,935	74.8	0.4
Reported excellent, very good, or good emotional health	10,336	75.5	0.4
Reported fair or poor emotional health	2,316	72.2	0.9

*The sample sizes in these categories do not sum to 12,652 because of nonresponse.
**The sample sizes in these categories do not sum to 12,652 because 95 individuals were in households that did not designate a financial respondent (an "R1"). Of those 95 individuals, 51.2 percent gave access to their Social Security records.

Table 2
Estimated Logit Model for Probability That
Social Security Earnings History Is Available

Explanatory Variable	Coefficient Estimate	Standard Error Estimate
Male	-0.182	0.045
Age	0.0057	0.0042
Born in U.S.	0.080	0.071
Nonwhite	-0.426	0.051
Completed high school	0.042	0.057
Some college	-0.123	0.067
Completed college	-0.120	0.073
Retired	0.019	0.069
1991 household earnings in $10,000	0.016	0.006
Never worked	-0.317	0.107
Any government employment	-0.010	0.084
Homeowner	-0.083	0.055
Household net worth in $100,000 (W)	-0.050	0.009
W squared	0.00078	0.00021
Excellent, very good, or good physical health	0.105	0.063
Job-limiting disability	0.129	0.061
Excellent, very good, or good emotional health	0.156	0.060
Sample size	12,469*	
Chi-square for model	164.6	

*183 observations are omitted from the logit analysis because of missing data, as described in the notes to Table 1.

References

Gustman, Alan L., and Thomas L. Steinmeier. 1999. "What People Don't Know about Their Pensions and Social Security: An Analysis Using Linked Data from the Health and Retirement Study." Working Paper No. 7368, National Bureau of Economic Research.

Juster, F. Thomas, and Richard Suzman. 1995. "An Overview of the Health and Retirement Study." *Journal of Human Resources* 30 (Supplement): S7-S56.

Olson, Janice A. 1999. "Linkages with Data from Social Security Administrative Records in the Health and Retirement Study." *Social Security Bulletin* 62 (2): 73-85.

Venti, Steven F., and David A. Wise. 1998. "Lifetime Earnings, Saving Choices, and Wealth at Retirement." Unpublished.