USING AUXILIARY INFORMATION TO ADJUST FOR NON-RESPONSE IN WEIGHTING A LINKED SAMPLE OF ADMINISTRATIVE RECORDS

Barry W. Johnson and Paul B. McMahon, Internal Revenue Service
Presented at the 2002 American Statistical Association

Federal estate tax returns are a rich source of information on the assets and liabilities associated with decedents, as well as data on beneficiaries of estates. When linked with income tax data for the decedents and their beneficiaries, the resulting data base provides a unique opportunity to study a variety of important economic issues relating to the transfer of wealth and the accumulation of capital. However, in creating such a complex, linked data base, it is inevitable that, for a variety of reasons, a number of records would be missing.

In this paper, we detail steps taken to weight the linked files. We adjust the linked record weights in two stages. First, an adjustment factor is created to balance to the original population totals, essentially treating unmatched records as non-respondents. Next, we employ auxiliary data, post-stratification, and raking to adjust the sampling weights and then compare those results to estimates from other administrative record sources.

Background

The Federal estate tax is a tax on the transfer of assets from a decedent's estate to its beneficiaries and is, therefore, levied on the estate. It is not an inheritance tax. The estate tax, the gift tax, and the generation-skipping transfer tax, together, form the Federal unified transfer tax system. This system taxes transfers made by individuals both during life and at death.

A Federal estate tax return, Form 706, must be filed for every U.S. decedent whose gross estate, valued on the date of death, combined with certain gifts made by the decedent, equals or exceeds the filing threshold applicable for the decedent's year of death. The return must be filed within 9 months of a decedent's death, unless a 6-month extension is requested and granted. All of a decedent's assets, as well as the decedent's share of jointly owned and community property assets, are included in the gross estate for tax purposes and reported on Form 706. Also reported are most life insurance proceeds, property over which the decedent possessed a general power of appointment, and certain transfers made during life.

Expenses and losses incurred in the administration of the estate, funeral costs, and the decedent's debts are allowed as deductions against the estate for the purpose of calculating the tax liability. A deduction is allowed for the full value of bequests to the surviving spouse, including bequests in which the spouse is given only a life interest, subject to certain restrictions. Bequests to qualified charities are also fully deductible.

The Statistics of Income Division (SOI) of the Internal Revenue Service selects a sample of Federal estate tax returns filed during the calendar year as part of its annual estate study. These data are used for budget analysis, tax law evaluation, and other economic studies. From time to time, a subsample of estate tax returns, collectively referred to as an "estate collation study," is selected for further processing.

The collation subsample is designed to collect additional data on decedents and the beneficiaries of their estates. Some of these data are drawn from Form 706 and supplemented with information provided in wills and trust documentation. Income tax data from Form 1040 for both decedents and beneficiaries are also linked to data from the Federal estate tax return. Bequest data, combined with income data for beneficiaries, can be used to study bequest patterns and motives (see Joulfaian, 1994), as well as to better understand the effects of inheritances on certain beneficiary behaviors (see Mikow and Berkowitz, 2000). Income tax data linked to estate tax data for decedents can be used to study such issues as the relationship between realized income and wealth (see Steuerle, 1985) and the usefulness of the life-cycle model of savings for explaining bequest behavior (see Modigliani, 1988).

The Data

The design for the 1992 Estate Collation Study had four main stages, starting with the selection of the Statistics of Income 1992 Estate Tax Return Study sample. This sample of Federal estate tax returns filed between 1992 and 1994, inclusive, was designed for use in estimating both tax revenues in all 3 calendar years and personal wealth holdings for 1992 decedents. The 3-year sample period was devised to ensure that nearly all returns filed for 1992 decedents would be subjected to sampling, given the long lag that can occur between a decedent's death and the filing of an estate

tax return, due to extensions.[1] The design had three stratification variables: size of total gross estate, age at death, and year of death. Total gross estate (the sum of all the asset valuations) was chosen as a stratifier to satisfy the first use, estimating tax revenue, and was limited to five categories:

- $600,000 under $1 million,
- $1 million under $2 million,
- $2 million under $5 million,
- $5 million under $10 million, and
- $10 million or more.

Age was selected as a stratifier, in part, because personal wealth estimation is based on death rates, which are closely correlated with age. The decedent's age at death was disaggregated into five categories: less than 40, 40 under 50, 50 under 65, 65 under 75, and 75 or older (including age unknown). The year-of-death variable was separated into two categories based on whether the year of death was 1992 or another year. This outline was designed in late 1990 and implemented in 1992, with minor sampling rate changes for non-1992 decedent strata in Calendar Years 1993 and 1994. The sampling probabilities for the 20 strata for 1992 decedent estates were not changed over the sampling period.

Estate tax returns were sampled during administrative processing, without regard to the possibility of any audit examination. A portion of the sample was selected because the decedents' Social Security number (SSN) ending digits corresponded with those in the Social Security Administration's Continuous Work History Sample (CWHS). However, the majority of returns were selected on a flow basis using a Bernoulli sampling method. The actual sampling mechanism creates a permanent random number based on an encryption of the SSN (see Harte, 1986). Sample rates were preset based on the desired sample size and an estimate of the population. They ranged from 3 percent to 100 percent, with more than half of the strata selected with certainty. These samples were limited to returns filed for decedents with total gross estates of at least $600,000, the estate tax filing threshold in effect for this period. Of the 28,530 returns sampled between 1992 and 1994, 11,943 were for decedents who died in 1992.

[1] An examination of returns filed between 1982 and 1992 revealed that almost 99 percent of all returns for decedents who die in a given year are filed by the end of the second calendar year following the year of death. Further, the decedent's age at death and the length of time between the decedent's date of death and the filing of an estate tax return are related (see Johnson, 1998). Therefore, it was possible to predict the percentage of unfiled returns, within age strata, and to adjust the final 1992 year-of-death sample weights to account for returns not filed by the end of the 3-year sampling period.

Collation Study Data

A subset of returns filed for decedents who died in 1992, and for whom an estate tax return was filed in either 1992 or 1993, was selected for inclusion in the 1992 Estate Collation Study. The subsample was limited to these 2 study years because of time restrictions for extracting the particular IRS Master File data in which we were interested. Because one study goal was to examine the relationship between income and wealth for decedents, it was necessary to have income data for, at minimum, the last full year prior to death. The source records on the Individual Master File (IMF) that we required were only retained for those posting in the current calendar year and the 2 immediately previous years (other types of records had longer retentions but contained insufficient data for our needs). Thus, in order to acquire Tax Year 1991 individual return filings (submitted in 1992), we had to cut off our selection for this collation study after the Calendar Year 1993 Form 706 selections. Estate tax returns filed during 1994 for decedents who died during 1992 had to be ignored in the sampling process. This truncation of the sample period, however, introduced significant bias since complex estate tax returns, especially those for large estates, take the most time to prepare. Much of the work documented in the rest of the paper focuses on trying to reduce the effects of this bias on estimates generated from the collation data base.

In focusing on returns filed for 1992 decedents, we eliminate 20 strata from the original estate study sample. The sample of 1992 decedents was itself further reduced from that of the original SOI sample of estate tax returns for several reasons. First, our sponsor, the Treasury Department's Office of Tax Analysis was primarily interested in the larger estates due to their expectation that only larger amounts passed to heirs would have a discernable impact on their behaviors. Second, some of the individual income tax return data were to be collected by taking advantage of the Statistics of Income Individual Program's panel selection procedures. This panel operation was an adjunct to the standard stratified Bernoulli sampling that is the mainstay of that series. There was, however, a limit on the number of SSN's that could be added to that operation due to hardware constraints. The subsample rates ranged from 4 percent to 100 percent. Returns that indicated that a decedent had made bequests to living beneficiaries, but

for which important bequest information was not reported, were rejected from the final data set.[2]

At that point, we had two sampling processes and one frame constraint that affected the sample. In addition, there was one other administrative issue that should be considered. Due to the way that SOI computer operations are planned, programmed, and tested, the sample rates are developed almost 18 months prior to implementation, based on desired sample size and filing projections that are developed using prior-year data. However, there was a recession in 1992, which diminished the value of many estates. Thus, our actual sample was smaller than expected, both for the basic estate study and the collation study. The final collation study sample contained 4,525 decedent records. These estates reported 22,000 beneficiaries, including some beneficiaries whose bequests were contingent on either the death or coming of age of other, more primary beneficiaries.

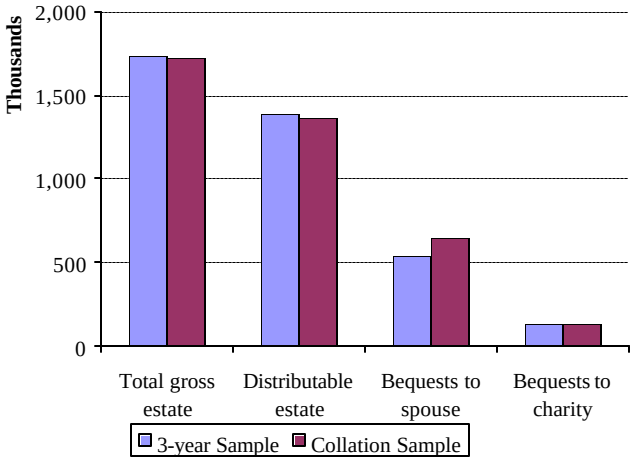
Base Weights

To calculate a base collation study weight, we needed to consider all the mechanisms that were actually involved in selecting the subsample. In order to account for the truncation of the sampling period, we post-stratified to the existing 3-year strata population counts. However, this did not fully address the reason that some returns are filed later than others. Discussions with estate tax practitioners revealed that returns reporting a significant tax liability take the longest to prepare, since several valuation experts may be consulted prior to determining final asset values, in order to minimize, as much as legally possible, the tax liability. Thus, to calculate a collation study sample weight, we further post-stratified on a binary variable indicating whether or not an estate had reported a tax liability. Note that, in both cases of post-stratification, we had the population from which the sample was drawn to tally for the strata totals. Figure A compares selected estimates using the final, weighted collation study decedent data with those from the full, weighted 3-year estate study file.

Decedent 1040 Files

For decedents in the 1992 Estate Collation Study, income tax data were obtained from the IMF for the tax period ending December 31, 1991, the last full year prior to a decedent's death. The data available were limited to those necessary for effective tax

Figure A: Mean Values for Selected Variables, 3-Year Sample vs. Collation Sample



administration. Income tax data were available for 3,767 of the 4,525 decedents in the collation sample, a linkage rate of 89.5 percent. Linkage rates varied substantially by sex and sample code. A decade ago, the IRS administrative records processing system for Individual Income Tax Returns did not include a provision for ensuring the quality of the secondary, or spousal, SSN. Since the spousal SSN on the tax form is usually that of the wife, we felt that an adjustment to the weights had to be made along gender lines to compensate for the higher level of non-matches. Indeed, while almost 90 percent of the returns filed for male decedents could be matched to a Tax Year 1991 Form 1040 return, the link rate for female decedents was only slightly more than 70 percent.

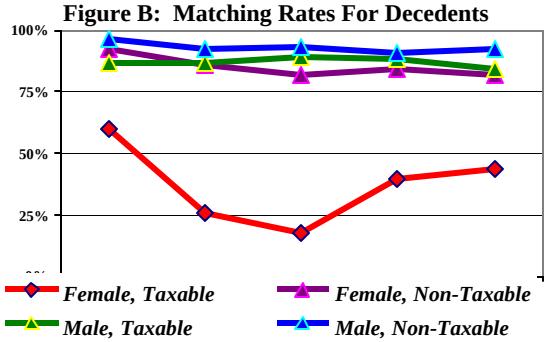
Refining this further, we found that 92 percent of the male decedents with taxable estates and 88 percent of the male, non-taxable estates were matched. This is not an important difference. Only slightly lower than those groups were the non-taxable estates of females. However, as Figure B shows, the largest difference was in the case of the taxable estates of female decedents, whose records had a match rate of only 35 percent. In fact, we were able to match only 18 percent of records for the estates of women with taxable estates valued between $2 million and $5 million. This is partly due to the very small samples in this category, which totaled only about 100 across the five size categories.

Weight adjustments for the matched 1040 returns were, thus, calculated within the original sample strata, post-stratified by gender and tax status. In several instances, samples were very small, making it necessary to collapse strata. Wherever possible, strata were

[2] In cases where a preparer had failed to provide beneficiary information on IRS Form 706, every attempt was made to collect this information from supplemental documentation, such as will and trusts. In the end, there were 22 returns that were rejected due to missing information.

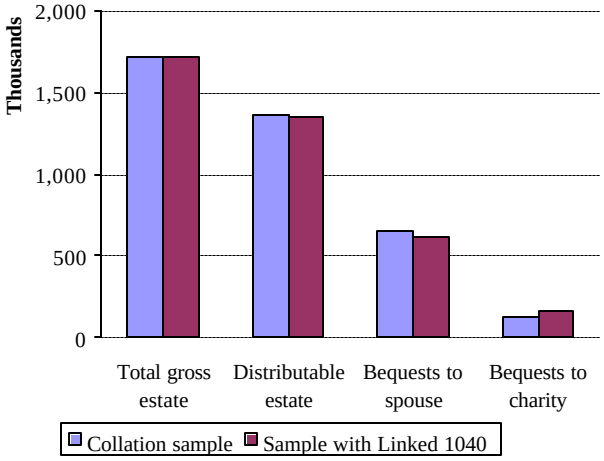
collapsed across tax status, rather than sample

Figure B: Matching Rates For Decedents



code, to preserve the original sample probabilities. In some cases, small samples required additional aggregation. The resulting adjustments were applied to the collation study base weights. Figure C compares selected estimates using the final, weighted collation study decedent data with those from the weight-adjusted linked 1040 file.[3]

Figure C: Mean Value for Selected Variables, Collation Decedent Sample vs. 1040 Linked File



Beneficiary 1040 Files

For the purposes of studying the income effects, if any, that arise from receiving an inheritance, it was necessary to collect data from a Form 1040 filed prior to receiving an inheritance, to use as a baseline, and similar return data reflecting income after the receipt of the inheritance. For the 1992 Estate Collation Study, we selected beneficiary income tax returns for tax periods ending in 1992 (baseline) and 1995 (to see any effects of inheritance).[4]

The 1040 data for collation study beneficiaries were collected from two sources. Data for 1992 came from the IRS IMF for returns filed during Calendar Years 1992 and 1993 with tax periods ending December 31, 1992, the year of our decedents' deaths. Data for tax periods ending December 31, 1995, came from returns filed during Calendar Years 1995 and 1996 and were collected as a part of the SOI Individual Income Tax data program. Linkages were initially based on SSN matches and were confirmed by comparing name information present on Form 706 with that on Form 1040. Contingent beneficiaries (those whose inheritances were conditioned on the death, coming of age, or disclaimer of another beneficiary) were not considered in this analysis.

There were 10,983 beneficiaries for whom income tax information was available for tax periods ending 1992 and 1995, a linkage rate of 55.1 percent, much lower than that of decedents. The actual linkage rates varied substantially by sample code. An adjustment similar to that calculated for the decedent 1040 data was indicated. However, in this case, there were additional possible explanations for non-matches. First, for some beneficiaries, the preparer may have refused to provide an SSN, since it is not used for tax administration purposes. Second, for beneficiaries whose bequests were in the form of a trust, the entity identification number (EIN) associated with the trust may have been reported instead of the beneficiary's SSN. Third, transcription errors introduced either during the preparation of the original return or during data collection were also possible. Additionally, some beneficiaries may have been too young to have ever filed income tax returns in one or both periods, while others who had filed in 1992 may have died before 1995. These last possibilities introduce some uncertainty as to the exact population of beneficiaries for whom Form 1040 data should have been available. The first step in calculating final weights for this file, then, was to determine the appropriate population to use in adjusting the base weights.

In determining the population of beneficiaries whom we believed should have filed a Form 1040 in both periods, it was necessary to know the age of each beneficiary. An individual's date of birth was available

[3] These estimates differ slightly from those in Figure A because they are limited to decedents who had made bequests to living beneficiaries. A small group of decedents selected into the collation study had limited their bequests to charitable organizations.

[4] 1992 was chosen over 1991 due to the availability of more complete data for that filing year. Because of delays associated with settling an estate, beneficiaries who received inheritances from 1992 decedents would not have received them in Calendar Year 1992.

from Social Security Administration (SSA) records and was automatically present for nearly all beneficiaries for whom a Form 1040 for either 1992 or 1995 was available. For the remaining beneficiaries for whom a seemingly valid SSN had been reported, we tried linking to an SSA file, known as the Data Master One (DM1) file, which contained dates of birth. Of the 8,940 non-matched beneficiaries, we were able to obtain a DM1 file match for 2,200. Thus, age was still missing for the 5,295 beneficiaries for whom no SSN had been reported, as well as for the 1,445 beneficiaries for whom a seemingly valid TIN had been reported, but for whom no linkage to either 1040 data or the DM1 file was possible.

An examination of the distribution of a few key variables suggested that there was no significant difference between the groups of beneficiaries for whom age was known and those for whom age was missing. In the absence of any systematic bias, it was possible to impute missing ages using the hotdeck imputation method (see Hinkins and Scheuren, 1986). Donor cells were created, based on a beneficiary's relationship to a decedent and the decedent's age. Beneficiary age and an indicator as to whether or not a beneficiary had died prior to 1995 were selected randomly with replacement from the donor cells. Once this was completed, an examination of the data suggested that a beneficiary who was at least 18 in 1992 could have reasonably been expected to file in both periods. Consequently, beneficiaries whose actual or imputed ages were less than 18 were dropped from the analysis, as were those who had died prior to 1995. These constraints reduced the original sample of 19,926 to 18,663 non-contingent beneficiaries of 1992 estates for whom 1040 data would have been expected.

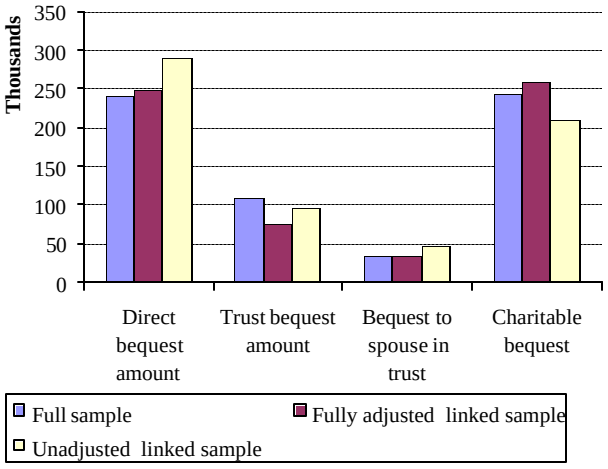
Initial weight adjustments were calculated within the original decedent sample code, thus preserving the original probabilities of selection, and were then post-stratified by tax status. The resulting initial weights were applied to the file, and weighted frequency estimates were generated by relationship to the decedent. Figure D shows that there were significant differences between the weighted estimates by relationship category for the full beneficiary sample and those produced from the linked sample. Thus, ratio raking was indicated. In addition to adjusting by relationship, we examined the possibility of separating the data further by tax status and gender. Further analysis, however, indicated that the decedent's sex was not related to the non-response bias; thus, only relationship and tax status were used. For some relationship categories, the sample was too small. So, these were combined with similar relationship categories for adjustment purposes. Adjustments were calculated and applied to the previously adjusted weights. The counts by sample code and tax status were then reproduced, using the now twice adjusted weights. Two more rounds of raking adjustments were made, each time adjusting first within the sample code

Figure D: Frequency Estimates Before Raking Adjustments

Relationship to decedent	Full sample estimate	Linked 1040 sample estimate	Percent under/over estimate
Surviving Spouse	27,023	35,274	30.5
Child	64,946	74,835	15.2
Grandchild	22,689	23,248	2.5
Sibling	11,449	9,156	-20.0
Niece/nephew	36,541	37,294	2.1
Parents	1,329	777	-41.5
Other relative	26,352	19,116	-27.5
Not related/unknown	26,953	14,823	-45.0
Total number beneficiaries	217,282	214,523	-1.3

and tax status and then within the collapsed relationship category. At this point, weighted frequency estimates from the matched 1040 file and the entire beneficiary file were nearly identical by both sample code and relationship category so that no more adjustments were indicated. Figure E compares selected estimates using the final, weighted collation study beneficiary data with those from the weight-adjusted linked 1040 file.

Figure E: Mean Value, Selected Variables, Full Beneficiary File vs. Linked 1040 File



Conclusion and Future Plans

While the 1992 Estate Collation Study data base has great research potential, biases, introduced by small sample sizes and non-response problems, provide significant challenges. Particularly troubling was the necessity of truncating the sampling period from 3 to 2 years in order to conform with administrative records processing systems. Adjusting the sample weights, using post-stratification and raking, seems to be a practical method of reducing some of these biases for particular types of analyses.

The work presented in this paper suggests several additional research projects. First, the estimate for bequests through trust from the beneficiary linked data file was significantly lower than the value estimated using the full beneficiary sample file (see Figure E). This bias was not surprising, given that, while only beneficiaries with an SSN were included in the linked data file, a trust EIN was very often reported instead of an SSN when a beneficiary's entire bequest was in the form of a trust. Additional post-stratification by the form of bequest might reduce this bias. Second, we would like to measure the variances of our estimates in order to test whether differences between the mean values calculated using the linked files with adjusted weights, and those produced using the larger estate tax samples, are significant. Calculating variances, however, will require significant resources given the relative difficulty of producing variance estimates for stratified and linked datasets. Third, the post-stratification results from this work suggest that the same approach could be used to improve estimates from the annual estate study samples, although more research will be needed to determine the appropriate post-stratification classes.

Future collation studies will be affected by a number of recent developments. SOI has already undertaken a collation study of 1998 decedents with a much larger sample size. Other developments, such as IRS efforts to improve the quality of secondary SSN's on the IRS Master File and a new SOI archive of IMF data for a long time-series of tax years, should reduce some of the most troubling sources of bias present in the 1992 collation study data base. Studies beyond that of 1998 decedents will be limited by recent legislative changes that increase the estate tax filing threshold incrementally for decedents who die between 1999 and 2009 and then eliminate the tax entirely for decedents who die after December 31, 2009.[5]

[5] The Economic Growth and Tax Relief Reconciliation Act of 2001 calls for the repeal of the estate tax for decedents dying after December 31, 2009. However, that legislation expires after December 31, 2010. It is unclear, at present, whether or not the repeal of the tax will be made permanent.

References

Harte, James M. (1986), "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *American Statistical Association 1986 Proceedings of the Section on Survey Research Methods*, pp. 603-608.

Hinkins, Susan and Scheuren, Frederick (1986), "Hotdeck Imputation Procedure Applied to a Double Sample Design," *Survey Methodology*, Volume 12, pp. 181-196.

Johnson, Barry W. (1998), "Updating Techniques for Estimating Wealth From Federal Estate Tax Returns," *American Statistical Association 1998 Proceedings of the Business and Economic Statistics Section*, pp. 143-147.

Joulfaian, D. (1994), "The Distribution and Division of Bequests in the U.S.: Evidence from the Collation Study, OTA Paper 71, U.S. Department of the Treasury.

Mikow, J. and Berkowitz, D. (2000), "Beyond Andrew Carnegie: Using A Linked Sample of Federal Income and Estate Tax Returns To Examine the Effects of Bequests on Beneficiary Behavior," *American Statistical Association 2000 Proceedings of the Section on Government Statistics and Section on Social Statistics*, pp. 183-188.

Modigliani, F. (1988), "The Role of Intergenerational Transfers and Lifecycle Savings in the Accumulation of Wealth," *Journal of Economic Perspectives* 2, pp. 15-40.

Steuerle, C. E. (1985), "Wealth, Realized Income, and the Measure of Well-Being," in *Horizontal Equity, Uncertainty, and Economic Well-Being*, M. David and T. Smeeding, editors, University of Chicago Press, Chicago.