Developing an Optimal Approach to Account for Late-Filed Returns in Population Estimates

Cynthia Belmonte, Brian Raub, Paul Arnsberger, Charles Day[1]
[1]IRS, Statistics of Income, 1111 Constitution Ave NW, Washington DC 20024

Abstract
Estimates for populations of interest for Statistics of Income (SOI) programs are produced by drawing stratified, random Bernoulli samples of tax and information returns as they are filed, over predetermined sampling periods that often span multiple years. While this methodology results in the inclusion of the majority of targeted returns, a small number of returns for each study are filed beyond the data collection period, potentially introducing non-response bias into the population estimates. For a given sampling period, the paper will analyze historical filing patterns to develop an approach for accounting for late-filed returns. This research will assess the weight adjustment approach currently used in SOI's estate tax study and will provide a basis for application of a similar approach in each of the exempt organizations and private foundations studies.

Key Words: non-response bias, population estimates, post-stratification, Bernoulli sampling

1. Data Sources and Background

The Statistics of Income (SOI) division of the Internal Revenue Service (IRS) collects and disseminates detailed data based on samples of administrative records, including tax and information returns. The SOI sampling frame for any given study consists of tax or information returns posted to the appropriate IRS return transactions processing system within a designated time period. Often, this time period is the statutory period within which taxpayers are required to file. For other studies, in which taxpayers may file returns over many years, sampling occurs over a designated time period in which past experience tells SOI statisticians all but a small fraction of returns will be filed. In either event, some taxpayers may file returns for the period of interest after sampling for a study has ended. Over the years, SOI has taken several approaches to adjusting for the potential incompleteness of its sampling frames, some on a case-by-case basis and others more uniform in nature.

Building on previous research, this paper describes three SOI studies covering tax and information returns for estates, private foundations, and exempt organizations and briefly outlines current practices for handling late-filed returns [1]. Next, the authors describe two models for predicting the proportion of late-filed Estate tax returns using several covariates. Using the 2004 year-of-death sample, the authors will then apply and evaluate the new adjustment factors by comparing results to known population totals and previous estimates derived using existing adjustment factors.

1.1 The Estate Tax Study

With its annual Estate Tax study, SOI extracts demographic, financial, and asset data from Federal estate tax returns. The annual study allows production of a data file for each filing, or calendar, year. By focusing on a single year of death for a period of 3 filing years, the study allows production of periodic year-of-death estimates. A single year of death is examined for 3 years, as over 98 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. Data included in this paper are for Year of Death 2004 and were obtained from returns filed in Calendar Years 2004-2006.

The estate of a decedent who, at death, owns assets valued in excess of the estate tax applicable exclusion amount, or filing threshold, must file a Federal estate tax return, *Form 706, U.S. Estate (and Generation-Skipping Transfer) Tax Return*. For decedents who died in 2004, the exclusion amount was $1.5 million. Alternate valuation may be elected only if the value of the estate, as well as the estate tax, is reduced between the date of death and the alternate date. The estate tax return is due 9 months from the date of the decedent's death, although a 6-month filing extension is allowed. In some cases, longer filing extensions may be permitted.

For the Year of Death 2004 Estate Tax study, there were 11,817 Form 706 returns in the sample selected from a population of 42,424. The SOI Estate Tax study is classified into strata based on year of death, the size of total gross estate, and age of the decedent. For the Year of Death 2004 study, there were a total of 57 sampling strata, with sampling rates ranging from 4 percent to 100 percent.

1.2 The Private Foundation and Exempt Organization Studies

The annual SOI studies of private foundations and exempt organizations collect detailed financial data, as well as information on charitable and grant-making activities and compliance with IRS regulations, from information returns filed by exempt organizations. Studies are conducted for a single tax year and include samples of returns filed and processed during the 2 calendar years immediately following the target tax year. Data discussed in this paper for the Private Foundation and Exempt Organization studies were obtained for Tax Year 2004 returns filed and processed to the IRS Business Masterfile during Calendar Years 2005 and 2006. While this 2-year sampling period ensures almost complete coverage of the target population, there are still a number of returns processed after the close of the second year (i.e., December 31, 2006 for the Tax Year 2004 study), which are generally excluded from the samples.

Private foundations and nonexempt charitable trusts are required to file Form 990-PF (*Return of Private Foundation or Section 4947(a)(1) Nonexempt Charitable Trust Treated as Private Foundation)* annually. Similarly, certain exempt organizations are required to file Forms 990 *(Return of Organization Exempt from Income Tax)* or Form 990-EZ *(Short Form Return of Organization Exempt from Income Tax)*. SOI conducts annual studies based on samples of Forms 990-PF, 990, and 990-EZ filed for a given tax year. These information returns are due 5 months after the close of the organization's accounting period, although a 3-month filing extension is allowed. In some cases, additional filing extensions may be granted.

For the Tax Year 2004 Private Foundation study, there were 7,805 Form 990-PF returns in the sample, selected from a population of 80,570. The SOI Private Foundation study is classified into strata based on the size of end-of-year fair market value of assets, with each stratum sampled at a different rate. Sampling rates ranged from 1 percent for private foundations with total assets less than $125,000 to 100 percent for private foundations with total assets of $10 million or more.

The Tax Year 2004 exempt organization sample of section 501(c)(3) filers comprised 15,070 Forms 990 and 990-EZ, selected from a population of 279,415. End-of-year book value of assets was the stratifying variable for the exempt organization study. Sampling rates ranged from 1 percent for exempt organizations with total assets less than $500,000, to 100 percent for those with total assets of $50 million or more.

2. Current Treatment of Late-Filed Returns

SOI's estate, private foundation, and exempt organization studies all share a common challenge in accounting for returns filed after the end of the designated sampling period. The Estate Tax study Year-of-Death estimates include weight adjustments for late-filed returns. Such adjustments were first developed in 1997 by Woodburn, and later updated in 2007 by Raub. Weight adjustment factors are calculated using historical data from the IRS Masterfile, and vary by size of estate, age of decedent, and tax status of return. The aim of using these weight adjustments is to improve the overall population estimates, as well as the estimates for the subpopulations of returns that have historically filed late with greater frequency. To the extent that late-filers create bias in the Estate tax estimates, this approach seems to be an effective strategy in mitigating this bias. Another strength of this approach is that the data used to calculate the adjustment factors are readily available in the IRS Masterfile.

In contrast to the estate tax study, population estimates for the private foundations study do not include standard adjustment factors to account for returns filed after the close of the 2-year sampling period. Instead, during file closeout, efforts are made to identify and include late-filed returns that would have been sampled at the 100-percent rate (i.e., organizations with fair market value of assets of $10 million or more). This allows for more complete coverage of the target population by including returns that would have been selected with certainty. This allows for time-series analysis of a specific organization (or panel of organizations). Potentially, this treatment can extend the two-year sampling period by 4 to 5 months, the typical length of time between the end of the normal sampling period (in December) and the creation of the final study file (in mid-May). This can introduce some inconsistency from year-to-year, since the slightest variation in the Master File processing cycle, file review schedule, or final delivery date can affect the sampling period from one year to the next. Additionally, this method does not specifically address smaller organizations, which account for the largest share of the late-filing population.

3. Methodology and Results

The goal of the current research is to determine whether the current estate tax study adjustment factors still accurately reflect taxpayer behavior. Additionally, the authors seek to develop and assess alternative methods of estimating adjustment factors on the estate tax study, and whether such methods can be applied to other studies (Private Foundations, Exempt Organizations) that are subject to similar late-filing challenges.

The authors propose adjusting the weights of the returns in the estate tax return sample by multiplying by the inverse of the predicted proportion of returns filed by the cutoff of sampling. In order to not overly inflate variance, it is desirable that a relatively small number of adjustments be applied to the returns. Rather than attempting to calculate an adjustment based on each return's values of selected covariates, the adjustment factors were calculated for specific categories that are either sampling strata, groups of strata, or subsets of a stratum. Such an adjustment accounts for returns that will be filed after the end of the sampling period for the estates of decedents who died during the reference year.

Discussions with the estate tax study analyst yielded three possible explanatory covariates: size of the estate (measured by the total gross estate value), age of the decedent, and taxability of the estate; that is, whether or not an estate tax was due before the application of credits. Taxability is naturally a categorical variable. While age is discrete, it can take on over 100 values, thus age categories, similar to the categories used in constructing sampling strata, were used as dummy variables, as were size categories. The categories were chosen to reflect marginal changes in late-filing behavior based on exploratory analysis. Precise category boundaries were then adjusted due to the desire to have them, when possible, match sampling stratum boundaries, and the need to have sufficient numbers of late-filing events in each cross classification (taxablilty × age × size) to support modeling.

3.1 Survival Analysis

Survival analysis, or time-to-event modeling, is a well-known technique for measuring the probability that some event (death in its original application) will occur within a given time period. Since its original application, it has been applied to model more general time-to-event problems. The survivor function estimates the probability of an event occurring at or after some time *t*. In the context of this research, the event of interest is the filing of an estate tax return, and "survival" equates to making it to the end of the sampling period cut-off (3 years) without filing an estate tax return.

One method for developing such a model is Proportional Hazards (Cox) regression. Cox regression is a widely accepted type of survival analysis model. It allows the use of covariates to help explain differences in times to some event for different observations. For the estate tax study, age of the decedent and size of the estate are both important predictors of time to filing. Cox regression can also handle other important features of the estate tax study data.

In order for an estate to come into existence, someone must die. Prior to his or her death, and the formation of the estate, there is no risk of an estate return's being filed. SOI conducts a study of estates of decedents who die in every third year. Since the dates of death are distributed throughout the reference year, estates are formed and become subject to filing at different times. This is similar to a study of cancer treatments, where subjects may enter the study at time of diagnosis, and thus, subjects may become part of the study cohort at different times. The phenomenon of some subjects' beginning to experience positive probability of an event's occurring at a later time than others is called "delayed entry," and the observations for those subjects are referred to as "left-truncated."

Using Cox regression, the authors estimated the parameters of the survivor function conditional on the values of the covariates. For every adjustment stratum (shown in Table 1), the authors fit a model to the estate tax study year-of-death 2001 population data. In order to do this, the authors analyzed all of the possible combinations of the selected covariates for each stratum, keeping the best set of significant covariates for each stratum. The authors also used the year-of-death 2004 sample file to create a vector of all three covariates for each return. The authors then used the covariate vectors from the 2004 sample to predict a set of survival probabilities.

Table 1: Definition of Categories of Total Gross Estate and Age

Variable Name	*Lower Bound*		*Covariate*		*Upper Bound*
ageCats0	0	≤	Age	<	40
ageCats1	40	≤	Age	<	65
ageCats2	65	≤	Age	<	70
ageCats3	70	≤	Age	<	75
ageCats4	75		Age		or older
sizeCats0	$1.5 million	≤	Total Gross Estate	<	$2.0 million
sizeCats1	$2.0 million	≤	Total Gross Estate	<	$3.0 million
sizeCats2	$3.0 million	≤	Total Gross Estate	<	$5.0 million
sizeCats3	$5.0 million	≤	Total Gross Estate	<	$10.0 million
sizeCats4	$10.0 million		Total Gross Estate		or more

3.1.1 Survival Analysis Results

Table 2 presents new population estimates derived using the survival analysis approach as well as comparisons to known population totals and estimates using previous adjustment methods. The survival analysis model overestimated number of returns by about 6.5 percent and total gross estate by 10 percent.

Table 2: Year-of-Death 2004 Population Totals and Sample Estimates with Adjustment Factors Modeled Using Survival Analysis

Weight Adjustment Method	*Number of Returns*	*Percentage Difference*[1]	*Total Gross Estate ($ Millions)*	*Percentage Difference*[1]
Population total	**41,922**	**n.a.**	**149,430**	**n.a.**
Unadjusted estimate	40,453	-3.50	147,163	-1.52
Woodburn (1992)	40,785	-2.71	148,199	-0.82
Raub (2007)	40,867	-2.52	148,502	-0.62
Belmonte *et al.* (2010)	44,680	6.58	163,942	9.71

[1] Percent difference from known population total

The overestimation of both number of returns and total gross estate indicate that non-proportional hazards were not ignorable. The models were fit with time-dependent covariates to adjust for the effect of time on the effects of the different covariates. Many of the time-dependent covariates were highly significant. Also, their associated hazard ratios were greater than one, indicating that hazard, or risk, of filing increased as time passed. By ignoring the violation of proportional hazards, the hazards across time were essentially “averaged over”. This led to an underestimation of hazard, resulting in survival probabilities for late-filed returns higher than acceptable for the desired outcome.

3.2 Logistic Regression

Filing before or after the designated sampling cutoff can be modeled as a binary response variable. Logistic regression is a commonly used method for predicting the proportion of times an event occurs in a number of trials conditional on the values of some explanatory covariates [2, 3]. As in the previous model, the selected covariates were size of the estate (again, measured by the total gross estate value), age of the decedent, and taxability of the estate. Definitions of the selected categories are shown in Table 3.

Table 3: Definition of Categories of Total Gross Estate and Age

Variable Name	*Lower Bound*		*Covariate*		*Upper Bound*
ageCats0	0	≤	Age	<	40
ageCats1	40	≤	Age	<	65
ageCats2	65	≤	Age	<	70
ageCats3	70	≤	Age	<	75
ageCats4	75		Age		or older
sizeCats0[1]	$2.0 million	≤	Total Gross Estate	<	$3.0 million
sizeCats1	$3.0 million	≤	Total Gross Estate	<	$5.0 million
sizeCats2	$5.0 million	≤	Total Gross Estate	<	$10.0 million
sizeCats3	$10.0 million		Total Gross Estate		or more

[1] In consideration of the potential application of this methodology for producing the year of death 2007 estimates, the lower bound for total gross estate used in this model reflects the exclusion amount in effect for decedents who died in 2007.

3.2.1 Logistic Regression Results

Table 4 shows the analysis of maximum likelihood estimates. All categories of all covariates are highly significant. Model development was guided in part by residual analysis, influence measures, and goodness-of-fit tests, but, as this paper is primarily concerned with good predictions and not explanation, these are omitted here.

Table 4: Analysis of Maximum Likelihood Estimates

Parameter[1]	*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > Chi-Sq*
Intercept	1	-2.4574	0.1316	348.8	< 0001
ageCats1	1	-0.5127	0.1311	15.3	< .0001
ageCats2	1	-0.7958	0.1385	33.0	< .0001
ageCats3	1	-0.9031	0.1356	44.3	< .0001
ageCats4	1	-1.3849	0.1286	116.0	< .0001
sizeCats1	1	-0.1191	0.0400	8.9	0.0029
sizeCats2	1	-0.2640	0.0525	25.3	< .0001
sizeCats3	1	-0.5813	0.0786	54.7	< .0001
Taxable	1	-0.1385	0.0413	11.2	0.0008

[1] The effect of the first category of each of the dummy variables for Age and Total Gross Estate is reflected in the Intercept.

Results from this method were quite good. Table 5 presents new population estimates derived using the logistic regression model as well as comparisons to known population totals and estimates using previous adjustment methods. This method produced an excellent estimate of total number of returns, the predicted value for which the method was designed. Additionally, the method resulted in a reasonable estimate of total gross estate.

Table 5: Year-of-Death 2004 Population Totals and Sample Estimates with Adjustment Factors Modeled Using Logistic Regression (for Returns with Total Gross Estate of $2.0 million and above)

Weight Adjustment Method	*Number of Returns*	*Percentage Difference*[1]	*Total Gross Estate ($ Millions)*	*Percentage Difference*[1]
Population total	**28,355**	**n.a.**	**161,007**	**n.a.**
Unadjusted estimate	27,701	-2.31	159,330	-1.04
Woodburn (1992)	27,926	-1.51	160,245	-0.47
Raub (2007)	27,981	-1.32	160,582	-0.26
Belmonte *et al.* (2010)	28,315	-0.14	162,213	0.75

[1] Percent difference from known population total

4. Future Steps

Estimates for the Estate Tax study benefit from a small adjustment to account for late-filed returns. As the research shows, logistic regression can be a useful method for calculating such adjustment factors. Results from logistic regression models are encouraging for the future development, assessment, and potential application of such models to adjust population estimates for other SOI studies. The authors recommend further evaluation of the new estate tax study adjustment factors using the available population data and estimates of other highly reliable variables. Additionally, the authors recommend development of similar models for each of the Private Foundation and Exempt Organization studies.

5. Acknowledgements

The authors would like to acknowledge the management of the Statistics of Income Division, Internal Revenue Service, for their support and encouragement of this work. The views expressed in this paper are those of the authors and should not be attributed to the Internal Revenue Service or the U.S. Department of the Treasury.

6. References

[1] Raub, B., C. Belmonte, P. Arnsberger, M. Ludlum. The Effect of Late-Filed Returns on Population Estimates: A Comparative Analysis. In *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA: American Statistical Association, 2009.

[2] Agresti, A. *An Introduction to Categorical Data Analysis*, John Wiley & Sons, New York, NY, 1996.

[3] Stokes, M., Davis, C., Koch, G. *Categorical Data Analysis Using the SAS System*, SAS Institute, Cary, NC, 1996.