CREATION OF PANEL DATA FROM CROSS-SECTIONAL SURVEYS

Stephanie Hughes and Susan Hinkins, Internal Revenue Service
S. Hughes, Statistics of Income CP:R:S:P, P.O. Box 2608, Washington, DC 20013-2608

Key Words: Longitudinal Data, Missing Data

Introduction

One of the IRS Statistics of Income (SOI) staff"s primary functions is to produce and publish annual estimates of various corporate tax data items. In order to do this, SOI collects and processes a sample of corporate tax returns for each tax year and creates a microdata file from which the estimates are obtained; this microdata file is also delivered to our primary users at Treasury's Office of Tax Analysis (OTA). Much of OTA's work concerns estimating revenue and modeling the effects of proposed policy. Because they need to model and estimate behavior of corporations, they require information on the behavior of a corporation over time. Therefore, a few years ago SOI and OTA formed a working group to see how they could best use their resources to create a panel file.

Both SOI and OTA have agreed all along that they do not want to sacrifice the precision of the cross-sectional estimates for the benefit of creating a panel file. Since OTA's work is policy driven, they cannot predict what properties or what types of corporations will be important for future modeling problems. Their needs often change depending on what issues are "hot" at the moment. OTA is looking for a panel that could be used by many different users for many different purposes.

SOI's corporate sample design, which is very good for producing annual estimates, also employs a simple sampling technique which results in as much longitudinal data as possible at no additional cost to the annual estimates. However, there are many missing data problems in the longitudinal data that are currently obtainable from the cross-sectional samples. In this paper, we will describe SOI's corporate cross-sectional sample design and the sampling technique which maximizes the amount of overlap from year to year, the proposed panel, and, finally, the missing data issues and their effect on the longitudinal data.

Cross-sectional Samples and Resulting Overlap

The sampling frame for the SOI samples consists of all returns that post to the Internal Revenue Service's (IRS) Business Master File (BMF). The annual sample is typically between 80,000 and 100,000 returns, and the associated population is approximately 4 million returns. It is a stratified probability sample, where the strata are based on the form type filed, which is related to the type of corporation (C-Corporation, S-Corporation, Life Insurance Company, etc.) and, within form type, the strata are based on the size of the corporation in terms of total assets and a measure of income. Because of the very skewed distribution of the population, Neyman allocation results in a wide range of sample rates which increase with the size of corporations in each stratum (i.e., the large returns are sampled with higher probability of selection than the small returns). Almost 20% of the sample consists of large corporations selected in take-all strata and the minimum sample rate is constrained to be no smaller than 0.25%.

Because the accounting period does not always coincide with the calendar year, a "year" of corporate tax data is defined in terms of the end of the accounting period. For example, the 1992 population of corporate tax returns is defined as all returns filed for an accounting period ending between July of 1992 and June of 1993. This is to ensure that the returns in the sample contain at least 6 months of the current year's tax data. Sample selection for each tax year occurs over a two year period, in order to cover the filing period for these returns. For example, the 1992 sample is selected between July 1992 and June 1994.

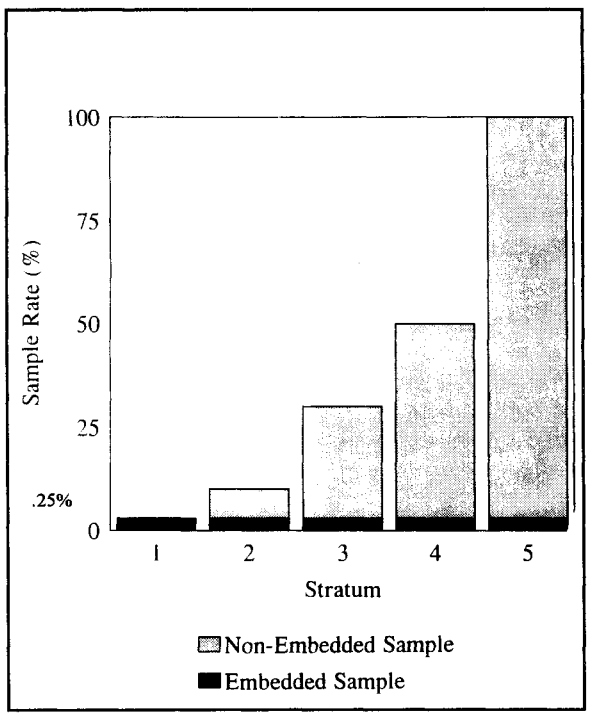
The selection technique uses each company's Employer Identification Number (EIN) to seed a pseudo random number generator, which generates numbers between 0 and 9999 with a uniform distribution. If the resulting random number falls under the respective stratum's sample rate times 10,000, then the return is selected for the sample. This method was first proposed and studied at the Bureau of Census by B. J. Tepping (1969). A more detailed description can be found in Harte (1986).

A large overlap occurs in the samples from year to year, because companies do not change their EINs often and the SOI sample design does not change significantly through the years. If an EIN is selected into the sample one year, it will be selected in all subsequent samples as long as returns for that EIN are filed each year and they fall into strata with the same or larger sample rates. Therefore, the overlap across samples consists primarily of large and either static or growing corporations.

Given the potential of being able to obtain a rich longitudinal file from the current samples, instead of concentrating on building a forward looking panel, the decision was made to try to create a panel file with the data that are currently available and to determine what

changes could be built in to improve it. For any time span of interest, one could create a panel file by matching companies in the respective cross-sectional files by their EINs. However, given the sampling technique, this overlap data can be considered as two separate pieces: embedded sample and non-embedded sample. Figure A is a pictorial representation of what we mean by the embedded and non-embedded sample members in the SOI samples.

Figure A.-- Embedded and Non-embedded Sample



The cross-sectional sample design is constrained to maintain a minimum sample rate of .0025. The embedded sample is made up of all corporations' returns that would be selected under this minimum sample rate (i.e., all EINs resulting in a random number < 25). This is a small random sample of all corporations in the population. These returns will be selected into the sample as long as they post on the Business Master File. Ideally, the only reason an embedded sample unit would no longer be selected into the sample is because it has died. Given that this is a random sample, the weighting issues are straight forward; however, the sample size is too small, particularly for the largest firms.

The non-embedded sample consists of all returns in the sample that would not be selected under the minimum sample rate (i.e., their random number ≥ 25). These units can potentially fall in and out of the samples between years due to decreases in sample rates or returns falling into strata with lower sample rates. This fact makes it difficult to identify the true births and deaths in the data and also makes weighting issues more complicated. However, this is a much richer sample, especially for the larger returns. Given the sample designs, there is a higher probability of having returns that either remain the same or increase in size with respect to their assets and/or income.

Proposed Panel

As mentioned earlier, the embedded sample has the advantage of being very straightforward to use, but it is not a sufficient sample for the largest firms. And while the non-embedded sample is a much richer sample, there are more complications involved in its use. Therefore, a proposal was made to construct a panel which contains pieces of both the embedded and non-embedded samples. This will be called the "combined panel."

For simplicity, first assume that there are no missing or incomplete data and that companies maintain the same EINs year to year. The population of companies over a given time span is comprised of two types of entities: companies existing in all years throughout the given time span and companies that are born or die during that time span.

The population of companies existing in all years throughout the time frame can be represented by all EINs present in the SOI samples every year during that time frame. This will be referred to as the overlap sample. The weight to be used for such a record is the maximum of the cross-sectional weights that were assigned to this company during those tax years. This applies to both the embedded and non-embedded sample units. The maximum weight is used so that these companies represent companies that were never in the sample, as well as companies that were in the sample for some but not all years within the given time frame.

The population of companies that are born or die during the time frame is represented by all the embedded sample EINs that are born or die during the time frame. These are identified by those embedded sample EINs that are not present in our samples every year. Weights of 400 are assigned to all of these corporations (400 is equal to the inverse of the minimum sample rate). Note that all non-embedded sample EINs that are not present in our samples every year are dropped from the panel.

Resulting Panel Data

The SOI cross-sectional files from tax years 1987 - 1992 were used to create a combined panel to examine the number of corporations in the overlap sample and in the embedded sample. There were approximately 35,000 companies in the overlap sample and an additional 5,000 companies in the embedded sample representing the births and deaths. The 40,000

companies in the combined panel give an estimated 8.3 million corporations for tax year 1992, which closely matches the cross-sectional estimate.

Figure B depicts some of the data patterns found in the combined panel. The figure does not represent the relative number of records with each pattern. The first row indicates the most common pattern, namely records in the combined panel with sample data for all six years.

Figure B.-- Potential Data Patterns Over Time

Pattern	< 1990	1990	1991	1992
1				
2			M	
3		M		
4		M	M	
5	?	?		
6	?	?	?	
7	?	?		?
8			?	?
9				?

M = Missing Return from Sample
? = Missing or Birth/Death

Patterns 2-9 show patterns of data found in the embedded panel with fewer than six years of data. These are the data that were intended to represent births and deaths. However, a surprising number of embedded sample members showed patterns of missing data, such as patterns 2-4. For example, if an embedded sample EIN is present in say 1989, 1990 and 1992, but not present in 1991, then this indicates missing data for 1991, since these units would be sample selected regardless of their size as long as they correctly file a return. Given the fact that we know there can be missing returns in the panel file, the patterns that look like births and/or deaths have to be questioned as well. We can no longer be sure that a return not available for a given year implies that the company did not exist that year. Therefore, we need additional information in order to correctly identify births and deaths.

The fact that there are missing returns among the embedded sample members also implies that there may be missing returns among the non-embedded sample members (i.e., there are other companies that did exist in 1987 - 1992, that we are not currently including in our combined panel because their returns are missing from any one of those years). If adjustments for the missing data are not made, then the resulting estimates may be biased.

Missing Data Issues

The extent of the incomplete data in the embedded panel means that we must be concerned with the impact that the missing returns have on the longitudinal data. It also points out how essential it is to be able to distinguish between missing data and the true births and deaths. However, one of the difficulties in this particular setting is that we do not always have a reliable indicator that a return is missing.

For the embedded sample EINs, there are two general causes of missing returns: nonresponse and noncoverage. Nonresponse is quite rare. There are routinely a few corporations that are sample selected but the returns are not available to SOI for editing. This may occur because the returns are in use by another IRS function (such as an audit) or are in a district office. It would be reasonable to assume that such returns are not missing at random. However, we can identify these returns, and we do have some data from the BMF for them. Therefore, missing data due to nonresponse could be dealt with through imputation or reweighting.

Noncoverage is the most common cause of missing data. The frame does not include all corporations of interest , primarily because corporations do not always file their tax returns in a timely fashion. Suppose a corporation is slightly late filing one year. For example, suppose the 1989 return was filed too late for the 1989 sampling process. In this case, both the 1989 and the 1990 return would have been in the 1990 frame. The 1990 return would be selected and the 1989 return would be rejected. This type of late filer can be fairly easily retained in the embedded sample. But there are also examples where the tax returns are six years late and, in effect , six years of data are filed simultaneously. How do we distinguish such a case of missing data from a death during the time before the returns are filed?

Another difficulty is really a problem of definition. A company can change it's EIN for various general reasons and the user will have to decide whether to treat it as the same company or treat it as a new company (i.e., the death of the old EIN and birth of the new EIN). Currently, SOI keeps track of the very largest returns in the population and ensures that those returns are included in the samples each year. In doing so, SOI does keep track of all EIN changes for these companies; however, SOI does not keep track of this for all returns in the population. For the purpose of this paper, therefore, we will assume that an EIN change results in a death and birth.

Adjustments for Missing Data

The major difficulty is that based solely on the information that is available in our samples, we can not conclusively distinguish between the missing data and the true births and deaths. We first consider the easier problem of just the embedded panel for tax years 1990 - 1992. In order to fill in some of the missing data patterns, we added the information for the rejected returns from the 1990 - 1992 samples. We also used the information we have for the embedded sample EINs prior to 1990, as depicted in Figure B.

We then used two extreme assumptions to distinguish the missing data from the births and deaths. Under the first assumption, records are assumed to be missing unless they have an initial or final return indicator (indicating a birth or death). Under the second assumption, records are assumed to represent births or deaths unless they are obviously missing from a particular year (i.e., missing between two years). For example, in Figure B, patterns 2-4 have missing data under either assumption. Under the first assumption, records with pattern 5 are considered incomplete unless there is a "first return" indicator on the record. Under the second assumption, pattern 5 is never considered incomplete but is always considered a birth. Figure C shows the differences between the two assumptions in estimating the number of companies identified as existing in all three years, versus births and/or deaths.

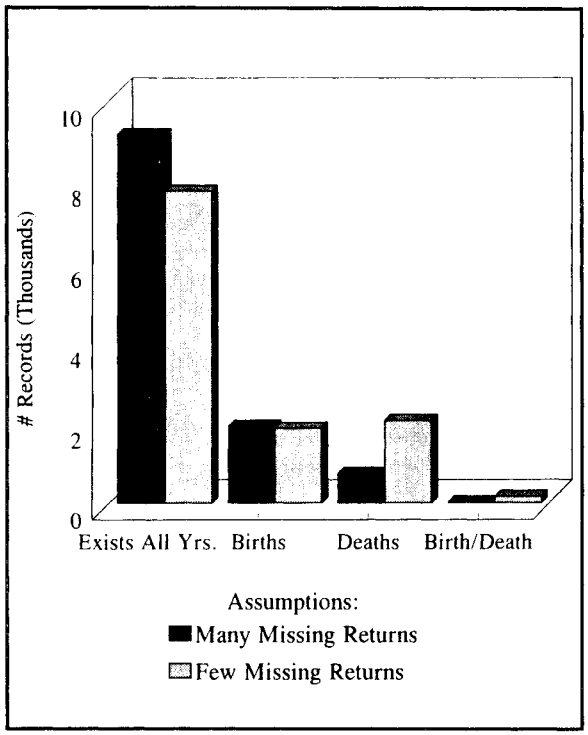
Since we have information for these EINs prior to 1990, but no information yet for these units after 1992, there is more information to determine births than deaths. We can see from the chart that there is little difference between the two assumptions in the estimated number of births. This would indicate that the initial return indicator may be quite reliable. The difference in the two assumptions is noticeable in the estimated number of deaths and in the estimated number of corporations existing in all three years.

A weighting adjustment was made to account for the missing records identified under each assumption. Forty-five weighting classes were defined, based on the company's years of existence, size of assets and net income. Most of the corporations identified as having missing returns are smaller corporations. The following weighting adjustment factor was calculated for each of these classes:

$$\text{Adj.} = \frac{(\#\text{ complete} + \#\text{ w/ missing returns})}{\#\text{ complete}}$$

The data do not appear to be missing at random. The adjustment factors range from 1.0 to 2.69, and the corporations having either no income or a loss appear most likely to have missing data.

Figure C.-- Embedded Sample Counts with Two Extreme Assumptions



The adjusted weights are equal to the original weight (400) multiplied by the resulting adjustment factors for each weighting class. The companies identified as having missing returns in any of the three years were dropped and the adjusted weights were applied to the remaining companies.

A comparison of the estimates showed that the estimates for deaths, in particular, had the most signifi-

Figure D.-- 1991 Estimates of Total Assets After Weighting Adjustments ($ in millions)

Assumptions for # Missing:

Pattern	Many	Few	% Diff.
Exist all 3 years	$11,154,655	$9,503,964	17.36
Births	309,590	310,016	-0.14
Deaths	55,025	227,660	-75.83
Birth & Death	9	4,063	-99.78
Totals	11,518,279	10,045,703	14.66

cant difference between the two assumptions. Figure D shows the difference due to the missing data assumptions in the estimate of the item Total Assets, by the pattern of data present.

The effects of the missing data assumptions on the estimates can be significant. Therefore, the treatment of the missing data in building the panel data requires careful consideration.

Conclusions and Future Work

Usually the modeling problems due to missing data are concerned with estimating properties of the data known to be missing. These are difficult modeling problems. In the situation described in this paper, there is an additional missing data problem. We must first address the difficulty of identifying which records have missing data versus records associated with births and/or deaths. Then we must model the missing data.

The two missing data assumptions compared above were about whether or not data were missing. The differences in the estimates were due to the differences in these two assumptions. The model for adjusting for the missing data was the same in each case. Therefore, it is very important to determine if there is more or better information regarding when an EIN is "born" or has "died."

Other sources of information on the population of corporations are also currently available to SOI. We are now in the process of determining the feasibility of using these additional databases to obtain information on the status of corporations in years they are missing from the SOI files. We are also considering how to best use the data available.

The additional databases include the following information:

- Population Name & EIN File, which includes the date of posting for the latest record filed;
- Parent and Subsidiary information on the population of consolidated filers; and
- Some information on mergers, bankruptcy status, filing extensions, EIN and name changes, carrybacks, etc.

In order to keep track of the status of corporations (i.e., missing, births, deaths) in the SOI samples, OTA and SOI have decided to create an Inventory File. This file would contain all corporations' Employer Identification Numbers that were present in any of the SOI samples back to tax year 1987. Each of those EINs, would have a status code for each year, which represents whether a return is available or, if not, the reason that the return is missing from the SOI sample (i.e., return not selected for that year, the company reorganized in some way and consequently obtained a new EIN, or the company did not exist that year).

This information can then be used in determining which companies should be included in a Combined Panel for any desired span of years back to 1987. For example, all the embedded sample companies that look like they had died (because they were not found in the SOI samples), but were found in the'Population Name & EIN File, with posting dates indicating they were still in existence will be treated as missing records and handled via weighting adjustments or imputation.

Even if the databases under consideration prove to be accessible and useful, there will always be the need to make some assumptions about whether certain units are missing or dead due to time delays in receiving information. However, we hope to minimize the need for such assumptions.

Given information or assumptions regarding which records include missing data, we must then investigate methods for compensating for the missing data, by reweighting or possibly using imputation when only one year's data are missing. With the current information, we are now beginning to consider how best to create a panel of 1987-1992 data.

Acknowledgments

A special thanks to Graham Kalton and David Morganstein, for their statistical expertise and advice; Paul Dobbins and Erik Larson, for creating the panel files; and Jeri Mulrow, for the research she had done on the other sources of corporate data available to SOI.

Bibliography

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

Hinkins, S., Mulrow, J.., Collins, R. (1990). Design and Use of an Imbedded Panel in the SOI Corporate Sample, *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 760-765.

Hinkins, S. and Scheuren, F. (1989). Design Modifications for the SOI Corporate Sample: Balancing Multiple Objectives, *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 654-658.

Hughes, S. (1994). Description of the Sample and Limitations of the Data, Statistics of Income...1991, *Corporation Income Tax Returns,* Publ. 16, 9-16.

Kalton, G. (1983). *Compensating for Missing Survey Data*, Research Report Series, Survey Research Center, Institute for Social Research, University of Michigan.

Mulrow, J., Hinkins, S., Shook, J. (1993). Statistics of Income Division's Uses of Administrative Business Tax Records: An Overview, *Proceedings of the International Conference on Establishment Surveys,* American Statistical Association, 28-36.

Tepping, B. J. (1969). Memo to Mr. J. F. Daly, Jan. 15, 1969, Department of Commerce, Bureau of the Census.