MODIFIED RAKING ESTIMATION IN THE CORPORATE SOI PROGRAM

Michael R. Leszcz, H. Lock Oh and Fritz J. Scheuren
Internal Revenue Service

This paper describes some modifications to the raking ratio estimation technique first proposed in a paper by Deming and Stephen in 1940 [1]. It discusses continuing research being conducted to improve the Corporation Statistics of Income (SOI) programs at the Internal Revenue Service. Variance and bias properties of the new estimates are examined, and some results from their application are provided.

The material is divided into five sections, as follows. Section 1 contains a brief background of the Corporate SOI program and previous SOI raking ratio estimation research. In Section 2 we describe the post-stratification raking ratio estimation technique, as applied to the Corporation SOI sample. Section 3 describes the new ideas incorporated into raking ratio estimation for this research, which include: bounding of the adjustment factors and exclusion of certain cells from the raking. Section 4 includes a description of the data file and data processing that were utilized in this research, and the results of the tests. In Section 5 the conclusions are summarized, and present and future applications of the estimators are discussed.

1. BACKGROUND

The Internal Revenue Service has produced statistics on the corporate tax returns annually since 1916. The major publications of these data include the annual publications, Statistics of Income Corporation Income Tax Returns [2], and the Source Book of Statistics of Income--Corporation Income Tax Returns [3]. The Source Book contains detailed tabulations of data available from SOI, and features income and balance sheet data classified by industry type and size of total assets. The broadest industry level in the source book is the twelve industrial divisions, and the lowest level is the 185 minor industries. Figure 1 represents an example of the hierarchical division of minor industries into major groups and industrial divisions. As can be seen in this representative example, Industrial Division 61, Wholesale Trade, is comprised of 3 Major Groups, 33 through 35, each of which is comprised of one or more minor industries. For example, Major Group 33 (Groceries and Related Products) contains 1 Minor Industry (5140--Groceries and Related Products), and Major Group 35 (Miscellaneous Wholesale Trade) contains several (e.g., 5010--Motor Vehicles and Automotive Products through 5190--Miscellaneous Nondurable Goods; Wholesale Trade Not Allocable).

FIGURE 1.--EXAMPLE OF INDUSTRY CODE GROUPINGS

Industrial Division	Major Group	Minor Industry
:		
61		
	33	
		5140
	34	
		5008
	35	
		5010
		:
		5190
62		
:		

The SOI data contained in these publications are utilized by various public sector organizations in the United States including the Treasury Department, the Congress, and the Commerce Department. They are also of use to various industrial planners and economists in the private sector.

The current research was preceded by a pilot study several years ago [4] in which it was indicated that post-stratification by major industry could result in large reductions in standard error for some items, with minimal increase in the standard error for others. The study utilized a stratification scheme in which tax returns were stratified by income, assets and industry--which is essentially the way the subject data file for this research is also stratified.

Figure 2 is a summary of the results of that pilot study. Note that the change in standard error ranges from a reduction of 16.9%, in the Inventories amount, to an increase of 1.0% in the Distribution to Stockholders amount. Also note that the Total Receipts amount indicates a reduction of 12.0% in standard error.

FIGURE 2.--REDUCTION IN THE STANDARD ERROR FROM POST STRATIFICATION BY INDUSTRY IN THE WESTAT PILOT STUDY

Item	Reduction In Std. Error (Percent)
Inventories	16.9
Business Receipts	12.4
Total Receipts	12.0
Base For Investment Credit	9.0
Depreciation	5.5
Taxable Income	3.9
Capital Gains	-.1
Distribution To Stockholders	-1.0

The returns that are the subject of our current research include all 1979 Corporation Tax Returns filed on Forms 1120 or 1120S which were sampled at less than 100%. As these two return types represent 98% of the corporate filers, we excluded from consideration for this research all

other 1979 Corporate return form types (e.g., 1120F, 1120L, 1120M, 1120DISC, and 1120P).

2. POST-STRATIFICATION RAKING RATIO ESTIMATION

The Corporate SOI sample is a stratified probability sample. The sample design calls for stratification of corporate tax returns by the size of assets and net income (or deficit) as defined in Figure 3.

FIGURE 3.--DETERMINATION OF SAMPLE STRATA

Size of Total Assets	Asset Code
$0 under $50,000	80
$50,000 under $100,000	81
$100,000 under $250,000	82
$250,000 under $500,000	83
$500,000 under $1,000,000	84
$1,000,000 under $2,500,000	85
$2,500,000 under $5,000,000	86
$5,000,000 under $10,000,000	87
$10,000,000 under $25,000,000 (financial)	88

Size of Net Income or Deficit (Absolute Value)	Income Code
$0 under $25,000	80
$25,000 under $50,000	81
$50,000 under $100,000	82
$100,000 under $250,000	83
$250,000 under $500,000	84
$500,000 under $1,000,000	85
$1,000,000 under $1,500,000	86
$1,500,000 under $2,500,000	87
$2,500,000 under $5,000,000 (financial)	88

Sample stratum are labelled 80 through 88. Each return receives an asset code and an income code. The return is assigned to a sample stratum corresponding to the higher of the two codes. Sample stratum 88 is limited to returns in the following financial industries: banks including mutual savings banks and bank holding companies, personal and business credit institutions, other insurance companies, and regulated investment companies.

Figure 4 presents the population counts and achieved sample size by strata for 1979.

The initial estimate of the Corporate SOI is obtained by inflating the sample count to the population control total by the categories (strata 80-88) used in the original sample design. The above approach we can call normal stratified sample estimation. Our interest in the present study is to improve an important class of estimates, those by industry, by incorporating an additional stratificaton of the population in the estimation process, after the sample has already been selected. This entails a two-way classification of the sample and a need to rake the sample count--not only to the population control by the original stratum, but also simultaneously, to the population control by the new stratum.

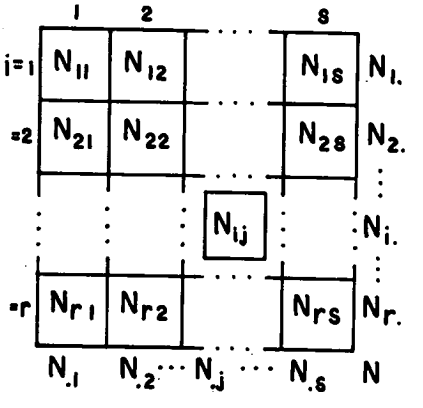
FIGURE 4.--CORPORATION POPULATION AND SAMPLE COUNTS BY SAMPLE STRATUM, 1979

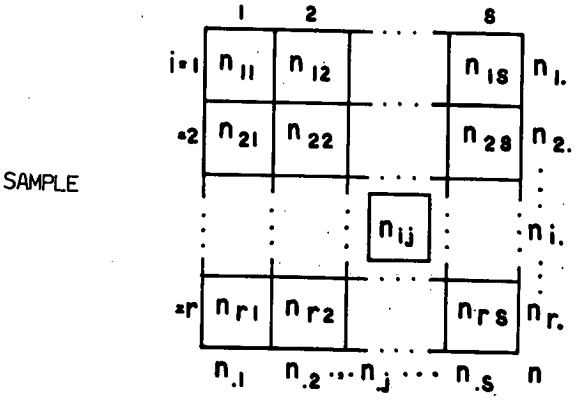
Stratum	Population	Sample Count
80	1,064,373	3,583
81	402,114	1,833
82	489,879	3,637
83	283,670	4,734
84	177,821	5,496
85	117,979	9,057
86	39,471	3,837
87	19,653	3,976
88	4,347	2,207
Other	49,839	41,708
TOTAL	2,649,146	80,068

The estimates, raking ratio estimates in the present case, are obtained by scaling iteratively the two-way table of sample counts to the two one-way population control (row & column) totals.

The idea of raking to the marginal totals is depicted graphically in Figure 5, as suggested by Deming [5]. In the universe, only the population marginal totals $N_{i.}$ and $N_{.j}$ are known--the individual cell counts N_{ij} are unknown. However, in the sample, all cell counts n_{ij} are known.

FIGURE 5.--GRAPHIC EXAMPLE OF RAKING TO MARGINAL TOTALS





The raking algorithm, described in the Appendix, involves successively adjusting sample counts by applying the ratios of row totals of population to sample, and then scaling the row adjusted sample counts in the preceding step to population column totals. This process is repeated until the adjustment ratios approach 1, within a tolerance.

3. MODIFICATIONS TO RAKING RATIO ESTIMATION

Under fairly general conditions, raking ratio estimation in contingency tables is entirely successful, i.e., optimal [6]. However, its use in adjusting sample weights is not always successful. One of the reasons for this is that if the variables used in the raking are not highly correlated with all of the variables in the sample, the weighting adjustments may lead to degradation in variance, rather than an improvement. Thus, it is a bit misleading to cite the fact that, for frequency estimates, raking yields best asymptotically normal estimation. In fact, for many variables in large scale surveys, raking can have a detrimental impact. This was our primary concern in the corporate sample. On the one hand, we hoped to gain significantly reduced variances for frequency estimates by industry. On the other hand, we were very desirous that only a minimal negative impact, if any at all, be made to variables not strongly related to industry. The schemes and their variants that we examined are summarized in Figure 6.

FIGURE 6.--ESTIMATORS CONSIDERED FOR THIS STUDY

Description	Abbreviation
National Weight (Normal Stratified Sample Estimation)	NORM
Pure Raking Ratio Estimation	PRRE
Bounded Raking Ratio Estimation	BRRE
Pure Raking Ratio Estimation (Excluding Cells with 200 or More Observations)	PRRE(200)
Bounded Raking Ratio Estimation (Excluding Cells with 200 or More Observations)	BRRE(200)
Pure Raking Ratio Estimation (Excluding Cells with 400 or More Observations)	PRRE(400)
Bounded Raking Ratio Estimation (Excluding Cells with 400 or More Observations)	BRRE(400)

First, we looked at cells where there were large numbers of sample observations. Since we knew the population totals for these cells, it was possible for us to use a simple ratio estimate in each such cell. The basic theory would show that such an estimation procedure would, for that single cell, be preferred to anything that could be obtained through raking [7]. What was not clear, was what the impact of taking that cell out of the population and out of the sample would be on the estimates in the other cells. The cell size limits that we utilized were 200 and 400. As we will see in the results, in some cases the total impact was beneficial, while in others not quite so beneficial.

The second variant was to constrain the raking adjustments so that they fell within a narrow range. The range used was between $\sqrt{2/3}$ and $\sqrt{3/2}$. This kind of constraint is often employed in simple ratio estimation [8]. The factors chosen are essentially heuristic, and we are not convinced by the work we have done so far that these factors are, in some sense, optimal, even for this population. More research is obviously going to be needed here. The approach to bounded raking ratio estimation is similar to that when large sample counts are available in a single cell. That is, it is similar in that, for the cell that is to be constrained, we bound the estimate; then take the estimated population total for that cell and the sample for that cell out of the population total and out of the sample; and then adjust the remaining observations. In the particular versions of raking that we employed, where large cells were excluded, we decided to rake to convergence for the remaining cells in the table. In the case of the bounded raking ratio estimation, we were a bit more conservative. We loosened the constraints on the other variables in the other cells in the table in hopes of speeding up the convergence. It is quite clear from the research that has been done on raking that tables with too many zeros in them will be very unstable and will not converge. The effect of both the procedures we are employing here is to introduce zeros into the table. If these zeros are strategically placed, or better, misplaced, then this can have a very serious detrimental impact on the rate of convergence and, even, on the quality of the estimators. Our recommendation before starting was, therefore, that the number of times that these procedures were employed would have to be fairly small. This is partly why we chose our limits for bounding ($\sqrt{2/3}$ and $\sqrt{3/2}$) and the maximum sample cell sizes of 200 and 400.

4. RESULTS

The root mean square errors (RMSE) of estimates for selected items were calculated for each estimator in Figure 6, utilizing the pseudo-replicate half-sample method [9].

The procedure involved: (1) construction of the half-samples; (2) two-way classification, sample code (original-stratum) by industry (post-stratum), of sample counts for each half-sample; (3) derivation of a set of weights for each sample for each estimator; (4) calculation of estimates of selected items by applying the weight to sample values for each sample; and (5) calculation of RMSE based on the variations of estimates each sample produced.

The resultant summary tabulations, presented in Figures 7 and 8, reveal what one would have expected of the frequencies. Substantial percentage reductions occurred for the PRRE, PRRE(200), and PRRE(400) estimates. Application of the bounding limits $\sqrt{2/3}$ and $\sqrt{3/2}$ decreased

FIGURE 7.
CORPORATE INCOME TAX RETURNS - 1979 STATISTICS OF INCOME
MEAN OF RATIOS OF ROOT MEAN SQUARE ERRORS BY TYPE OF ESTIMATOR AND SELECTED ITEMS RELATIVE TO PRRE

ESTIMATOR	NUMBER OF RETURNS			AMOUNTS					
	ALL RETURNS	WITH NET INCOME	FORM 1120S	TOTAL ASSETS	TOTAL RECEIPTS	NET INCOME	INCOME TAX	INVESTMENT CREDIT	JOBS CREDIT
NORM.....	71.10	1.70	1.23	1.10	1.09	.96	1.00	1.07	.97
PRRE.....	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BRRE.....	18.46	.99	1.00	.91	.94	.96	.94	.97	.96
PRRE(200).	1.00	1.00	1.00	.95	.96	.98	.98	1.00	1.00
BRRE(200).	19.70	1.03	1.00	.87	.90	.95	.93	.97	.96
PRRE(400).	1.00	1.00	1.00	.97	.99	.98	1.00	1.00	1.00
BRRE(400).	19.91	.99	1.00	.87	.92	.94	.92	.97	.96

Source: Calculated by averaging over all 58 major industry groups the ratios of the root mean square errors.

the magnitude of these reductions; however, they were still substantial. As the figures also indicate, in the Total Receipts amount tabulations we have similar, but less dramatic, changes when we consider each variant. The largest percentage reduction for this item occurred with the BRRE(200) estimator, while the least percentage reduction is evident with the PRRE estimator.

FIGURE 8.--MEAN OF RATIOS OF ROOT MEAN SQUARE ERRORS BY TYPE OF ESTIMATOR AND SELECTED ITEMS AS A PERCENT OF NORMAL STRATIFIED WEIGHTING ESTIMATES

Estimator	All Returns	Total Receipts	Jobs Credit
	I. Mean of Ratios		
PRRE.....	1.41	91.74	103.09
BRRE.....	25.96	86.24	98.97
PRRE(200)..	1.41	88.07	103.09
BRRE(200)..	27.71	82.57	98.97
PRRE(400)..	1.41	90.83	103.09
BRRE(400)..	26.60	84.40	98.97
	II. Percent Reduction		
PRRE.....	98.59	8.26	-3.09
BRRE.....	74.04	13.76	1.03
PRRE(200)..	98.59	11.93	-3.09
BRRE(200)..	72.29	17.43	1.03
PRRE(400)..	98.59	9.17	-3.09
BRRE(400)..	73.40	15.60	1.03

Source: See Figure 7.

It should be noted that, for Total Receipts, the decrease of 13.76% in the root mean square error, from the initial (NORM) estimate to that utilizing bounded raking ratio estimation, compares favorably with the 12.0% reduction in standard error that was experienced in the pilot study previously mentioned.

When considering the Jobs Credit amount tabulations, we see less dramatic changes. Included in some cases are (minimal) increases in the root mean square errors for this item, due to the fact that this field is less dependent upon the Industry Code grouping utilized in this research.

5. CONCLUSIONS AND FUTURE APPLICATIONS

We have shown in our research that the raking ratio variants considered here do, as speculated, result in reductions in the root mean square error for some items, while having a minimal adverse impact on others.

We have in fact, utilized the BRRE(200) variant considered in this research in our estimates of the 1980 and 1981 Corporation SOI data. Other SOI applications are under consideration; however, more research remains to be completed. Some of the questions to be considered in the future research include:

1) the degree to which we should rake to convergence;
2) optimal bounding factors that should be utilized;
3) optimization of the cell size limitation to be utilized;
4) application of bounding dependent upon the magnitude of the adjustment factor [10];
5) improvement of methods of variance estimation [11]; and,
6) consideration of the degree of instability that we create as we exclude more cells by the cell size limitation.

REFERENCES

[1] Deming, W.E., and F.F. Stephen, On a Least Square Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, Annals Mathematical Statistics, Vol. 11, pp. 427-444, 1940.

[2] U.S. Dept. of Treasury, 1980 Corporation Income Tax Returns, Internal Revenue Service, Publication 16, 1983.

[3] U.S. Dept. of Treasury, 1980 Source Book of Statistics of Income--Corporation Income Tax Returns, Internal Revenue Service, 1983.

[4] Westat, Inc., Results of a Study to Improve Sampling Efficiency of Statistics of Corporation Income, Bethesda, Maryland, January, 1974, (Unpublished).

[5] Deming, W.E., Statistical Adjustment of Data, Dover Publishing, New York, 1943.

[6] Ireland, C.T., and S. Kullback, Contingency Tables with Given Marginals, Biometrika, Vol. 55, pp. 1979-188, 1968.

[7] Oh, H.L., and F. J. Scheuren, Weighting Adjustments for Unit Nonresponse, Incomplete Data: The Theory of Current Practice, National Academy of Sciences, Panel on Incomplete Data (In Press).

[8] Hanson, Robert H., The Current Population Survey: Design and Methodology, U.S. Bureau of the Census, Technical Paper 40, 1978.

[9] McCarthy, P., Replication: An Approach to the Analysis of Data from Complex Surveys, Vital and Health Statistics, Public Health Service, Publication 1000, Series 2, No. 14, 1966.

[10] Suggestion by Roderick Little during the post-presentation discussion at the session in Toronto.

[11] Bankier, Michael D., Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys, 1983 Proceedings, American Statistical Association, Survey Research Methods Section.

APPENDIX--DEFINITION OF THE RAKING ALGORITHIM AND NOTATION

Notation

Population control total and sample, where

N_i = stratified (by sample code) population control total used in the original sample selection

N_{ij} = post-stratified (by sample code and PBA) population control total

n_{ij} = sample counts for $(i, j)^{th}$ stratum.

Initial estimate:

$\tilde{N}_{ij} = W_i n_{ij}$

$W_i = (\sum_j N_{ij})/(\sum_j n_{ij})$ = initial weight.

Post-stratified raking ratio estimate (PRRE):

$\tilde{N}_{ij}^{(1)} = w_{ij}^{(1)} n_{ij} = F_{ij}^{(1)} W_i n_{ij}$ = PRRE at the end of one iteration

$w_{ij}^{(1)} = a_i^{(1)} b_j^{(1)}$ = PRRE weight, where

$n_{ij}^{(1)} = \{(\sum_i N_{ij})/(\sum_i n_{ij})\} n_{ij} = a_i^{(1)} n_{ij}$

$n_{ij}^{(2)} = \{(\sum_j N_{ij})/(\sum_j n_{ij}^{(1)})\} n_{ij}^{(1)}$

$= b_j^{(1)} n_{ij}^{(1)} = b_j^{(1)} a_i^{(1)} n_{ij}$

$= \tilde{N}_{ij}^{(1)}$

$F_{ij}^{(1)} = w_{ij}^{(1)} / W_i$ = Post-stratified raking ratio adjustment (PRRA) factor

$\tilde{N}_{ij} = w_{ij} n_{ij} = F_{ij} W_i n_{ij}$ = PRRE at the end of convergence

$w_{ij} = \tilde{a}_i \tilde{b}_j$ = PRRE weight

$\tilde{a}_i = \prod_c a_i^{(c)}$ and $\tilde{b}_j = \prod_c b_j^{(c)}$

$F_{ij} = w_{ij}/W_i$ = PRRA factor ;

Post-stratified raking ratio estimate (PRRE) excluding cells with 200 or more sample count:

$\tilde{N}_{ij}^{*} = w_{ij}^{*} n_{ij} = F_{ij}^{*} W_i n_{ij}$

$$w_{ij}^{*} = \begin{cases} \tilde{a}_i \tilde{b}_j & \text{for } n_{ij} < 200 \text{ PRRE weight} \\ N_{ij}/n_{ij} & \text{for } n_{ij} \geq 200 \end{cases}$$

$F_{ij}^{*} = w_{ij}^{*} / W_i$ = PRRA factor.

Bounded post-stratified raking ratio estimate (BRRE) excluding cells with 200 or more sample count:

$\tilde{N}_{ij}^{**} = w_{ij}^{**} n_{ij} = F_{ij}^{**} W_i n_{ij}$

The BRRE weight w_{ij}^{**} is obtained by

1. For $n_{ij} \geq 200$

$w_{ij}^{**} = N_{ij} / n_{ij}$

2. For $n_{ij} < 200$

$w_{ij}^{**} = (1/K_i) w_{ij}^{*}$ for $\sqrt{2/3} \leq F_{ij}^{*} < \sqrt{3/2}$

$= (1/K_i) \sqrt{2/3} W_i$ for $F_{ij}^{*} < \sqrt{2/3}$

$= (1/K_i) \sqrt{3/2} W_i$ for $F_{ij}^{*} \geq \sqrt{3/2}$

where K_i is a norming constant such that for $n_{ij} < 200$

$K_i = (\sum_j N_{ij})/K_i \sum_j w_{ij}^{**} n_{ij}$

The PRRA factor is defined as

$F_{ij}^{**} = w_{ij}^{**} / W_i$.