Selected Semi-Variance Estimators of Underreporting Nonfarm Sole Proprietor Income

Chih-Chin Ho and William W. S. Chen, Internal Revenue Service

Underreporting nonfarm sole proprietor income on individual income tax returns has become a significant tax compliance issue. According to a recently published report by the Internal Revenue Service (IRS, 1996), less than 70 percent of net income earned by unincorporated businesses is correctly reported to the IRS. The study also shows that underreporting sole proprietor income accounts for $16.9 billion of the individual income tax gap in Tax Year (TY) 1992.

Current tax laws impose no limits on the amount of net business losses from a nonfarm sole proprietor (cases where total deductions exceed total receipts) for offsetting taxable income. As a result, underreporting nonfarm sole proprietor income is a linear combination of two types of reporting noncompliance: underreporting income-generating items (i.e., business receipts) and overstating income-offsetting items (i.e., business expenses). From this perspective, the IRS is interested in detecting both underreported receipts (UR) and overstated expenses (OE).

■ General Framework

From a standpoint of noncompliance detection, if we allow for the possibility that underreported receipts or overstated expenses are not symmetrically distributed, then a full variance estimator (FVE) of these compliance measures may not sufficiently capture variations within the less compliant segments and, therefore, undermine the "downside risk" in detecting both types of reporting noncompliance.

Full variance considers extremely high and extremely low underreported receipts or overstated expenses equally undesirable. Semi-variance, on the other hand, measures deviations from the mean for observations below or above the mean. As a result, a semi-variance estimator (SVE) can enable us to focus on selected segments of reporting noncompliance, such as those who engage in more aggressive UR (below the mean) and those who engage in more aggressive OE (above the mean). As a result, a semi-variance estimator would provide more realistic assessment of "downside risk" for detecting both UR and OE.

In this paper, we consider a set of selected semi-variance estimators developed in Josephy and Aczel (1993). We then apply these estimators to the selected reporting noncompliance measures pertaining to nonfarm sole proprietor income from the Taxpayer Compliance Measurement Program (TCMP) survey data.

Our statistical tabulations of the data show that the distribution of underreported receipts (UR) is skewed toward the left, and the distribution of overstated expenses (OE) is skewed toward the right. As a result, we define a lower variance estimator (LVE) for underreported receipts and an upper variance estimator (UVE) for overstated expenses, based on the semi-variance estimators presented in Josephy and Aczel (1993).

Since underreporting nonfarm sole proprietor income is a linear combination of UR and OE, the correlation between the more aggressive segment of UR and the more aggressive segment of OE is of considerable importance in understanding the interactive nature of these two types of reporting noncompliance.

To gain insight into this perspective, we extend the semi-variance concept to a covariance context and develop a semi-variance-based correlation coefficient (SCOR) to a joint distribution of a lower half of underreported receipts (UR_{LH}) and an upper half of overstated expenses (OE_{UH}).

■ Data Sources

The TCMP 1988 filer survey data (technically referred to as Phase III, Cycle 10) consist of 54,088 stratified random samples of approximately 104 million individual income tax returns for TY 1988. We use 11,132

returns with Schedule C (Profit or Loss from Business) for analyzing underreporting nonfarm sole proprietor income.

For these returns, both taxpayer-reported values and examiner-determined values are available for various line items relating to business receipts and expenses. As a result, the difference between reported and audited values of gross business receipts is calculated as underreported receipts (UR), while the difference between reported and audited values of total business deductions is calculated as overstated expenses (OE).

■ Semi-Variance Estimator

Based on Josephy and Aczel (1993), we consider a lower variance estimator (LVE) and an upper variance estimator (UVE):

$$LVE_x = (n/(n-1)^2) \cdot \sum(x_i-\bar{x})^2 \text{ for } x_i \le \bar{x} \quad [1]$$

$$UVE_x = (n/(n-1)^2) \cdot \sum(x_i-\bar{x})^2 \text{ for } x_i > \bar{x} \quad [2]$$

$$\text{where } \bar{x} = \sum\frac{x_i}{n} \text{ for all } x_i$$

The LVE (UVE) consists of the sum of squared sample derivations from the sample mean for those observations below (above) the sample mean, scaled by a factor to assure asymptotic unbiasedness.

■ Semi-Correlation Coefficient

We extend the semi-variance estimators described above to a covariance context and develop a semi-variance-based correlation coefficient (SCOR) between a pair of a selected half of X and a selected half of Y as:

$$SCOR(Hx,Hy) = [(n/(n-1)^2)\cdot\sum(x_i-\bar{x})\cdot(y_i-\bar{y})]/[\sqrt{SVE_x}\sqrt{SVE_y}]$$

$$\text{for } (x_i,y_i) \in S \text{ such that } x_i \in H_x \wedge y_i \in H_y \quad [3]$$

$$\text{where } \bar{x}=\sum\frac{x_i}{n} \text{ for all } x_i \qquad \bar{y}=\sum\frac{y_i}{n} \text{ for all } y_i$$

The selected semi-variance estimator (SVE) corresponds to the selected half of the variable: LVE_x for a lower half of X(LHx) and UVE_y for an upper half of Y (UHy).

The SCOR is based on a joint distribution of a selected half of variable X and a selected half of variable Y. It measures the correlation between two selected subsets of their respective variables. It is scaled by a factor to assure asymptotic unbiasedness.

■ Findings

Semi-Variance Estimates

We stratify 11,132 cases into 15 strata by taxpayer-reported values of business income to estimate the LVE and FVE for underreported receipts and the UVE and FVE for overstated expenses.

Table 1 presents the lower standard deviation (LSD) and full standard deviation (FSD) for underreported receipts. Table 2 presents the upper standard deviation (USD) and FSD for overstated expenses.

Audited Value-Based Stratification

In this subsection, we control for the values of examiner-determined values of business receipts or business expenses so that variations in true tax liability can be separated from variations in reporting noncompliance.

We stratify the entire 11,132 cases into 15 strata by examiner-determined value of business receipts to estimate the LVE and FVE for underreported receipts. Table 3 presents the estimates.

Similarly, we stratify 11,132 cases into 15 strata by examiner-determined values of business expenses to estimate the UVE and FVE for overstated expenses. Table 4 presents the estimates.

Semi-Correlation Coefficient Estimates

We select a lower half of underreported receipts (UR_{LH}) and an upper half of overstated expenses (OE_{UH}) as our selected subset S (UR_{LH},OE_{UH}). We estimate the semi-correlation coefficient (SCOR) for the joint distribution of UR_{LH} and OE_{UH} for 15 strata based on tax-

payer-reported business income. Table 5 presents the estimates.

To gain a clear perspective on how estimates of FCOR and SCOR differ, we also estimate FCOR for our full sample F (UR, OE) for 15 strata based on taxpayer-reported business income. Table 5 presents the estimates.

Balanced Bootstrap Replications

In this subsection, we use the bootstrap resampling method to test the stability of SCOR and FCOR estimation. For each of 15 strata based on taxpayer-reported business income, we create a set of 100 balanced bootstrap replications and calculate SCOR for the selected subset and FCOR for the full sample.

The method used to select balanced bootstrap samples was introduced by Davison, Hinkley, and Schechtman (1986) and refined in Hall (1992). Table 6 presents the averages for these estimates based on 100 balanced bootstrap replications.

■ Conclusion

Based on Josephy and Azcel (1993), we develop two semi-variance estimators as measures of "downside risk" in detecting reporting noncompliance on nonfarm sole proprietor income: a lower variance estimator (LVE) for underreporting business receipts (UR) and an upper variance estimator (UVE) for overstating business expenses (OE).

Table 1 and Table 2 show that both LVE for UR and UVE for OE are greater than their respective FVE counterparts. These findings are consistent with our prior notion that the distribution of UR is skewed toward the left, and the distribution of OE is skewed toward the right.

Table 3 and Table 4 show that results based on examiner-determined value-based stratification are very similar to those based on taxpayer-reported value-based stratification.

These findings illustrate the usefulness of using semi-variance measures in comparing distributions of UR or OE associated with various market segments. There are many instances where two distributions have similar FCV, but substantially different LCV or UCV. These semi-variance-based estimates provide alternative measures of "downside risk" in reporting noncompliance.

We extend the semi-variance concept to a covariance context. We develop a semi-variance-based correlation coefficient (SCOR) between the two least compliant groups in reporting business income: individuals who engage in both more aggressive receipts under-reporting and more aggressive expenses overreporting (UR_{LH}, OE_{UH}).

Table 5 shows that our SCOR estimates are negative for all 15 strata. These findings reflect a complementary relationship between the extent of receipts underreporting and the extent of expenses overreporting. In other words, for these "hard-core" noncompliant taxpayers, these two types of reporting noncompliance tend to reinforce one another.

Table 5 also presents our estimates for the FCOR based on the full sample (UR,OE) for all 15 strata. Except for two strata, these FCOR estimates are positive. These findings indicate a substitutive relationship between the extent of receipts underreporting and the extent of expenses overreporting. In other words, for general taxpayers at large, these two types of reporting noncompliance tend to substitute for one another.

Table 6 shows that average FCOR and SCOR estimates based on the 100 balanced bootstrap replications are very similar to results based on the original samples. These findings reflect the stability in estimating SCOR in spite of the relative small size of the selected samples (UR_{LH} , OE_{UH}).

■ Future Research

We would like to extend the SCOR estimation to other combinations of selected subsets. For example, it

would be interesting to know the SCOR between two lesser noncompliant reporting groups: the joint distribution of an upper half of underreported receipts (UR_{UH}) and a lower half of overstated expenses (OE_{UH}).

Furthermore, we would also like to explore the correlation between a lower half of underreported receipts and a full sample of overstated expenses (UR_{LH}, OE) or the correlation between an upper half of overstated expenses and a full sample of underreported receipts (UR, OE_{UH}).

■ Acknowledgments

The authors wish to thank Dennis Cox for his review and William Wong for his valuable suggestions.

■ References

Josephy, H. and Aczel, A. (1993), "A Statistically Optimal Estimator of Semivariance" in *European Journal of Operational Research*, Vol. 67, pp. 267-271.

Divison, A; Hinkley, D.; and Schechtman, E. (1986), "Efficient Bootstrap Simulation" in *Biometrika*, Vol. 73, pp. 555-556.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, (New York: Spring-Verlag).

Internal Revenue Service (1996), *Federal Tax Compliance Research: Individual Income Tax Gap Estimates for 1985, 1988, and 1992* (Publication 1415).

Table 1
Full Variance and Lower Variance Estimates of Underreported Receipts
by Taxpayer- Reported Business Income

Taxpayer-Reported Business Income		Full Standard Deviation (FSD)	Lower Standard Deviation (LSD)	Full Sample Mean	Full Sample Size	Lower Sample size
I1	-5,000 or Less	47,028	60,080	-5,884	629	165
I2	-5,000 - 0	20,092	33,201	-2,513	623	154
I3	0 - 2,500	18,227	43,024	-3,409	1,085	185
I4	2,500 - 5,000	11,595	26,393	-2,816	1,014	183
I5	5,000 - 7,500	14,046	33,027	-3,009	977	167
I6	7,500 - 10,000	15,292	38,164	-3,455	813	125
I7	10,000 - 12,500	21,902	56,539	-3,479	696	102
I8	12,500 - 15,000	20,913	50,123	-2,971	626	96
I9	15,000 - 20,000	21,107	54,022	-3,308	1,069	159
I10	20,000- 30,000	15,752	30,263	-1,773	1,169	188
I11	30,000 - 40,000	13,047	33,262	-2,242	654	92
I12	40,000 - 50,000	11,999	22,798	-1,857	372	65
I13	50,000 - 75,000	13,590	28,230	-1,671	565	79
I14	75,000 - 100,000	15,553	35,656	-2,367	339	48
I15	100,000 or More	36,100	101,588	-1,599	492	45

Table 2
Full Variance and Upper Variance Estimates of Overstated Expenses
by Taxpayer-Reported Business Income

TaxpayerReported Business Income		Full Standard Deviation (FSD)	Upper Standard Deviation (USD)	Full Sample Mean	Full Sample Size	Upper Sample Size
I1	-5,000 or Less	31,961	61,045	7,251	629	153
I2	-5,000 - 0	9,868	17,243	3,352	623	177
I3	0 - 2,500	7,382	13,200	2,132	1,085	290
I4	2,500 - 5,000	12,853	23,660	2,352	1,014	282
I5	5,000 - 7,500	6,109	6,688	1,775	977	329
I6	7,500 - 10,000	5,264	7,142	2,011	813	270
I7	10,000 - 12,500	10,375	13,469	2,130	696	219
I8	12,500 - 15,000	6,840	9,658	2,310	626	207
I9	15,000 - 20,000	6,324	9,302	2,223	1,069	331
I10	20,000- 30,000	11,165	12,377	2,774	1,169	358
I11	30,000 - 40,000	8,004	14,007	3,427	654	187
I12	40,000 - 50,000	11,607	13,729	2,824	372	117
I13	50,000 - 75,000	12,540	23,596	3,985	565	150
I14	75,000 - 100,000	11,892	15,312	3,693	339	102
I15	100,000 or More	10,946	14,639	4,113	492	144

Table 3
Full Variance and Lower Variance Estimates of Underreported Receipts
by Examiner-Determined Business Receipts

Examiner-Determined Business Receipts		Full Standard Deviation (FSD)	Lower Standard Deviation (LSD)	Full Sample Mean	Full Sample Size	Lower Sample Size
R1	5,000 or Less	31,919	4,970	4,938	628	576
R2	5,000 - 10,000	1,599	2,342	-275	597	101
R3	10,000 - 17,000	8,389	4,096	-169	660	161
R4	17,000 - 22,000	3,360	4,106	-230	853	121
R5	22,000 - 25,000	1,648	3,523	-471	864	160
R6	25,000 - 35,000	6,370	6,405	-1466	959	266
R7	35,000 - 50,000	5,288	10,382	-2155	847	183
R8	50,000 - 75,000	14,678	19,617	-2324	858	172
R9	75,000 - 100,000	11,052	22,310	-4066	542	116
R10	100,000 - 125,000	14,678	18,402	-1853	879	175
R11	125,000 - 150,000	13,379	27,620	-4193	705	152
R12	150,000 - 200,000	16,547	30,093	-4262	803	175
R13	200,000 - 300,000	24,417	54,408	-7263	848	158
R14	300,000 - 600,000	40,103	80,747	-9462	719	138
R15	600,000 or More	64,520	160,096	-16,499	362	55

Table 4
Full Variance and Upper Variance Estimates of Overstated Expenses
by Examiner-Determined Business Expenses

Examiner-Determined Business Expenses		Full Standard Deviation (FSD)	Upper Standard Deviation (USD)	Full Sample Mean	Full Sample Size	Upper Sample Size
E1	600 or Less	21,816	50,690	4,268	843	154
E2	600 - 2,500	10,805	22,873	1,959	764	169
E3	2,500 - 4,000	3,044	5,272	1,469	621	177
E4	4,000 - 5,500	2,804	4,435	1,519	578	189
E5	5,500 - 7,500	3,911	6,693	1,824	724	214
E6	7,500 - 10,000	3,788	6,111	1,853	719	236
E7	10,000 - 12,500	5,796	9,931	2,033	578	178
E8	12,500 - 16,000	15,860	30,766	2,975	679	177
E9	16,000 - 20,000	5,342	8,409	2,099	581	190
E10	20,000 - 30,000	5,299	7,974	2,282	945	314
E11	30,000 - 42,000	6,450	9,434	3,043	812	272
E12	42,000 - 60,000	11,122	17,074	3,337	882	299
E13	60,000 - 90,000	9,588	15,370	4,261	906	282
E14	90,000 - 180,000	12,721	15,046	3,983	996	352
E15	180,000 or More	29,458	34,377	5,504	496	177

Table 5
Full and Semi Correlation Coefficient Estimates of the Selected Joint Distributions of
Underreported Receipts and Overstated Expenses
by Taxpayer-Reported Business Income

Taxpayer-Reported Business Income		FCOR for (UR,OE)	SCOR for (UR_{LB},OE_{UB})	Full Sample Size (UR,OE)	Selected Sample Size (UR_{LB},OE_{UB})
I1	-5,000 or Less	0.42368	-0.52530	629	37
I2	-5,000 - 0	0.46709	-0.30185	623	35
I3	0 - 2,500	0.08052	-0.55617	1,085	65
I4	2,500 - 5,000	-0.10838	-0.26007	1,014	70
I5	5,000 - 7,500	0.40134	-0.49936	977	59
I6	7,500 - 10,000	0.06366	-0.63338	813	45
I7	10,000 - 12,500	0.30567	-0.52815	696	41
I8	12,500 - 15,000	0.39032	-0.24294	626	33
I9	15,000 - 20,000	0.08437	-0.27997	1,069	63
I10	20,000- 30,000	0.33963	-0.42016	1,169	82
I11	30,000 - 40,000	-0.14359	-0.33275	654	35
I12	40,000 - 50,000	0.38112	-0.42408	372	23
I13	50,000 - 75,000	0.44912	-0.70954	565	26
I14	75,000 - 100,000	0.28491	-0.59077	339	26
I15	100,000 or More	0.21221	-0.67802	492	20

Table 6
Average Full and Semi Correlation Coefficient Estimates of the Selected Joint Distributions of
Underreported Receipts and Overstated Expenses
Based on 100 Balanced Bootstrap Replications
by Taxpayer-Reported Business Income

Taxpayer-Reported Business Income		FCOR for (UR,OE)	SCOR for (UR_{LB},OE_{UB})
I1	-5,000 or Less	0.41420	-0.53477
I2	-5,000 - 0	0.49960	-0.37362
I3	0 - 2,500	0.08287	-0.52348
I4	2,500 - 5,000	-0.07667	-0.33098
I5	5,000 - 7,500	0.38344	-0.52350
I6	7,500 - 10,000	0.06075	-0.64051
I7	10,000 - 12,500	0.29780	-0.52673
I8	12,500 - 15,000	0.35558	-0.29722
I9	15,000 - 20,000	0.09335	-0.38396
I10	20,000- 30,000	0.26031	-0.42523
I11	30,000 - 40,000	-0.13711	-0.43368
I12	40,000 - 50,000	0.34741	-0.45619
I13	50,000 - 75,000	0.38655	-0.68677
I14	75,000 - 100,000	0.21587	-0.59546
I15	100,000 or More	0.18530	-0.63844