EVALUATING THE EFFECT OF SAMPLE SIZE CHANGES ON SCORING SYSTEM PERFORMANCE USING BOOTSTRAPS AND RANDOM SAMPLES

William Wong and Chih-Chin Ho, Internal Revenue Service

Presented at the 2002 American Statistical Association

Currently, the U. S. Internal Revenue Service (IRS) calculates a scoring formula for each return and uses it as one criterion to determine which returns to audit. Periodically, IRS updates this formula from a stratified random audit sample. In 1988, such an audit sample was selected. The sample was used to derive a new scoring formula. This score is one of the criteria used to determine whom to audit. The question was raised as to what size sample should be selected for the next audit sample. To answer that question, we wish to examine the effect of increasing or decreasing the sample by 20 percent has on the scoring formula. A very large audit sample would yield a scoring formula that would both increase the amount of revenue obtained from audits and decrease the burden of auditing those who filed accurately. But too large an audit sample would be self-defeating since we would be selecting many returns for the audit sample that would not result in more revenue from the sample and would increase the audit burden on those selected. No one likes to be audited, especially when an accurate return is filed.

Before evaluating the effect of audit sample reduction several problems had to be resolved. Both the scoring formulas used by IRS and the derivation procedures are confidential. Even treating it as a "black box" and running replications against it proved to be both sensitive and tedious. Instead, this paper chooses to analyze several simulated discriminant analysis methods of deriving a scoring function. The variance of this procedure is then calculated, using random samples and bootstrap samples. This analysis is repeated on new sample sizes, one 25-percent larger and one 20-percent smaller. For each of these samples, scoring functions are developed, scores are applied, and performance estimates are calculated. Finally, results across the discrimination methods and the three sample sizes are compared, using bootstrap and random sample estimation methods.

In the next section, we discuss our basic discriminant analysis methodology. To calculate average sample values and their variances, we use two basic types of procedures. We also outline the procedures used to generate random samples and those used to generate four different sets of balanced bootstraps. The results of our analysis are then presented, with the associated tables in the appendix. We also highlight our conclusions and future research and list references.

Discriminant Analysis Framework

We study one examination class with a sample of 4,356 audited returns. For our study purposes, we selected 100 original variables and use SAS Proc Stepdisc to determine which variables to use to create our discriminant function. Thus, the 100 variables are fixed, but the resulting subset of variables changes from sample to sample. We use a cross-validation approach to evaluate the performances of the scoring formulas.

In both random sample and bootstrap replicate methods, we start by selecting stratified samples using three strata. The weighted samples are first processed through SAS Proc Stepdisc to determine which subset of variables will be used. The classification variable used is a zero-one indicator of whether a return exceeds a minimum threshold discrepancy betwe en the reported and audited tax amounts. (Due to disclosure sensitivity, the threshold dollar amount is withheld.)

The weighted samples are then processed through SAS Proc Discrim using only the variables identified by the Proc Stepdisc procedure. Only parametric discrimination is tested. These weighted samples serve as the discrimination training data set. The discrimination test data set varies with the method tested. Since the discrimination test data set should not intersect with the training data set, the test data set is usually taken from the residual sample. The only exception is the Self Bootstrap, where we intentionally use the training data set as the test data set to determine the resulting bias.

One output of Proc Discrim is the posterior probability of the test return exceeding the threshold. This posterior probability is used as the score. The test data set returns are sorted by descending scores, and a cutoff percentage, *c*, of returns is selected for evaluation. The evaluation statistic is the "hit rate," which is defined as the portion of the selected weighted returns achieving the threshold. Cutoff percentages of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, and 75 are analyzed. The cutoff percentage of 100 is also tabulated to provide the average hit rates for the entire test sample.

Random Sample Framework

From our original sample of 4,356, we select our "large" stratified random subsample of 2,500 returns. We then select our "medium" stratified random subsample of 2,000 returns from the 2,500. Next, we select our "small" stratified random subsample of 1,500 returns from the 2,000. For each of our 400 Random Samples of the three sizes, we repeat this procedure. Each of these 400 Random Samples then serves as training data sets for our discrimination procedure. Each of the Random Samples is processed through Proc Stepdisc, using stepwise with p=0.15, to obtain optimum lists of variables by random sample to use in the Proc Discrim step. For the analysis, the untouched residual of 1,856 (= 4,356 - 2,500) returns serves as the Proc Discrim test data set. Note that the residual varies from sample to sample. Also, in order to compare results across sample sizes, the same residual is used as a test data set. Thus, the test data set for the i^{th} sample of size 2,500 is the same as that of the i^{th} sample of size 2,000 and 1,600.

Bootstrap Replicate Framework

We use the balanced bootstrap methodology suggested by Davison, Hinkley, and Schechtman (1986). In general, we obtain K balanced bootstrap samples from a sample X as follows:

1. Create K copies of X. Thus, if X had n units, K copies will have Kn units.
2. Randomize the Kn units.
3. Select the first n units for bootstrap 1. Select the next n units for bootstrap 2. Continue selecting until you have selected the Kth n units for bootstrap K.

These bootstrap samples are balanced in the sense that, across the sum of all bootstraps, every unit occurs exactly K times.

From our original sample of 4,356, we select our "large" stratified random subsample of 2,500 returns. We then select our "medium" stratified random subsample of 2,000 returns from the 2,500. Next, we select our "small" stratified random subsample of 1,500 returns from the 2,000. From each of the three subsamples, we then create 400 balanced bootstraps by applying the balanced bootstrap methodology described above to each of the three strata. Each of these 400 bootstrap samples then serves as training data sets for our discrimination procedure.

For the first bootstrap discrimination method, the Basic Bootstrap, we take each of our bootstrap samples and apply Proc Stepdisc, using stepwise with p=0.15, to obtain optimum lists of variables for the Proc Discrim step. For this analysis, the untouched residual 1,856 (= 4,356 - 2,500) returns serve as the Proc Discrim test data set.

For the second bootstrap discrimination method, the Forward Bootstrap, we proceed in a similar fashion to the Basic Bootstrap, except that we use forward discrimination with a maximum of 15 variables in the Proc Stepdisc step.

For the third bootstrap discrimination method, the Self Bootstrap, we proceed in a similar fashion to the Basic Bootstrap, except that we use the corresponding original random sample of size 1,600, 2,000, or 2,500 from which we bootstrapped as the test data set. Again, the purpose of this is solely to provide a measure of bias when using training data sets as test data sets.

The fourth bootstrap discrimination method, the Random Bootstrap, is a combination of the Random Sample method and the Basic Bootstrap. Here, we start by creating 400 Random Samples for each sample size as we did in the Random Sample Framework. We then apply the Basic Bootstrap assignments of frequencies to each (bootstrap, return) pair. For example, suppose we wanted the frequency to apply to the 8^{th} Random Sample, 21^{st} return for medium size samples. We obtain the medium size sample Basic Bootstrap frequency of the 8^{th} bootstrap, 21^{st} return. Note that, due to randomization in assigning returns to both bootstraps and random samples, the 21^{st} Basic Bootstrap return is very unlikely to be the 21^{st} Random Sample return. This method is an attempt to bridge the gap between the bootstrap results and the random method results. To prevent the "self test" bias, the test data set is the same as the one for the Random Sample method.

Results

For each of the methods, the mean hit rates across the 400 samples are calculated for each of the sample sizes by the percentage cutoffs. Along with each mean hit rate, the standard deviation of the mean is also calculated. These are tabulated in the Appendix.

Comparing the Basic Bootstrap with the Forward Bootstrap for a sample size of 2,500 indicates that the forward 15 variable bootstraps yield higher average hit rates for cutoff percentages under 9 percent. For cutoffs over 9 percent, stepwise with a p=0.15 is superior. Comparing the Basic Bootstrap with the Self Bootstrap shows that applying discrimination back to the training data set can greatly exaggerate the perceived performance. For 1-percent cutoff, we obtained 65 percent instead of 25 percent. For 10 percent, we obtained 32 percent instead of 22 percent. For every cutoff percentage, there is a clear positive bias. These results can be found in Table 1.

In both the Basic Bootstrap and the Random Sample methods, larger sample sizes resulted in larger hit rates. However, the rates are only marginally larger, and, sometimes, the difference is not significant. The sizable increase in sample size, from 1,600 to 2,500, yields very small increases in hit rates. These results can be found in Tables 2 and 3.

Comparing the Basic Bootstrap with the Random Sample for a sample size of 2,500 indicates that random samples have higher hit rates for cutoff percentages of less than 20 percent. These results can be found in Table 4.

Since the Random Sample estimates are true sample estimates, it appears that the Basic Bootstrap estimate has a negative bias for these cutoffs. But is this a true negative bias or is it a fluke of bootstrapping from just three samples and using just one test data set? We attempt to resolve this by computing Random Bootstraps. Random Bootstraps randomize both the three samples and the test data set by setting them to those used in the Random Sample method. Table 4 shows that the results are in the middle. While comparing the Basic Bootstrap method to the Random Sample method across all three sample sizes, we noticed that the Basic Bootstrap hit rates for sample size 2,500 were often just slightly below the Random Sample hit rates for sample sizes of 1,600. This is shown in Table 5.

Could it be that bootstrap replication or any replication of sample returns is ignored in discrimination procedures? On average, how many unique returns are there in a bootstrap? It turns out that for the bootstrap sample size of 2,500, the number of unique returns per bootstrap ranges from 1,539 to 1,624 with a mean of 1,582. This appears to confirm our suspicions.

Are there some inherent limitations with using bootstraps or any replication method to estimate discriminant properties? On reflection, there are estimates that replication methods obviously cannot estimate. Take the example of trying to estimate the number of unique (non-duplicate) returns in a data set. But our original task was to determine the relative increase or decrease in performance of a scoring function as we increase or decrease the sample. Since the proportion of unique returns is expected to remain constant across the sample sizes, the relative increase or decrease in performance should be preserved.

In general, are the differences between methods and sample sizes significant? The answer is predominantly yes. What about normality? According to the Shapiro-Wilk test, Basic Bootstraps were not normally distributed for cutoff percentages of 5 percent or less. Almost all the Random Samples did not fail the normality test. The Shapiro-Wilk test results are given in Table 6.

Conclusions

- Increasing the sample size from 1,600 to 2,500 returns yields rather minimal improvements in discriminant performance.
- Bootstrap estimates of hit rates appear to be negatively biased.
- Using the training data set as the test data set can greatly exaggerate the perceived performance.
- Forward discrimination using 15 variables appears to be mildly superior to stepwise with p=0.15 for small cutoff percentages.

Future Research

In the future we would like to test different forms of nonparametric discrimination and different ways of combining variables.

One possibility we would like to explore is what efficiency gain can we achieve by adding back variables based on different threshold-dependent scores? Preliminary work appears to indicate a potential gain.

Another technique we would like to try is to use discrimination to create a set of score variables from schedule-based discrepancies between the audit and taxpayer amounts. We would then add these variables to our list of the best 100 variables prior to running Proc Stepdisc.

References

Chernick, Michael R. (1999), *Bootstrap Methods, A Practitioner's Guide*, Wiley Interscience

Davidson, A.C., Hinkley, D.V., and Schechtman, E. (1986), Efficient bootstrap simulation, Biometrika 73, pp. 555-566

Davidson, A.C. and Hinkley, D.V. (1997), *Bootstrap Methods and their Application*, Cambridge University Press.

Efron, Bradley and Tibshirani, Robert J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall.

Hall, Peter (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.

Hjorth, J.S.Urban, (1994), Computer Intensive Statistical Methods, Validation, Model Selection, and Bootstrap, Chapman and Hall.

Appendix

Table 1-Comparing Average Hit Rates (AHR) and Std Dev (AHR) by Discriminant Method for Sample Size = 2,500

	Average Hit Rate (AHR)			Standard Deviation of AHR		
Cutoff %	Basic Bootstrap	Forward Bootstrap	Self Bootstrap	Basic Bootstrap	Forward Bootstrap	Self Bootstrap
1	24.94	25.71	64.97	0.40	0.41	0.35
2	25.22	27.09	54.35	0.27	0.30	0.25
3	25.65	27.14	47.92	0.22	0.24	0.20
4	25.33	26.30	43.57	0.18	0.20	0.15
5	24.63	25.25	40.48	0.16	0.17	0.13
6	24.01	24.45	38.21	0.13	0.15	0.11
7	23.46	23.79	36.22	0.12	0.13	0.10
8	23.01	23.14	34.66	0.11	0.12	0.09
9	22.60	22.60	33.29	0.11	0.11	0.08
10	22.31	22.11	32.03	0.10	0.10	0.08
15	20.75	20.30	27.23	0.08	0.08	0.06
20	19.70	19.24	24.28	0.07	0.06	0.05
25	18.88	18.39	22.25	0.06	0.05	0.04
30	18.14	17.64	20.68	0.05	0.05	0.04
35	17.45	16.95	19.42	0.05	0.05	0.03
40	16.76	16.30	18.32	0.04	0.04	0.03
45	16.15	15.78	17.38	0.04	0.04	0.03
50	15.58	15.30	16.60	0.03	0.04	0.02
75	13.23	13.30	13.66	0.02	0.02	0.01
100	11.81	11.81	11.62	0.00	0.00	0.00

Table 2--Comparing Average Hit Rates (AHR) and Std Dev (AHR) by Sample Size for Basic Bootstraps

	Average Hit Rate (AHR)			Standard Deviation of AHR		
	Sample Size			Sample Size		
Cutoff %	1,600	2,000	2,500	1,600	2,000	2,500
1	26.04	24.67	24.94	0.37	0.41	0.40
2	25.36	25.16	25.22	0.25	0.31	0.27
3	25.11	25.48	25.65	0.21	0.24	0.22
4	24.77	25.18	25.33	0.18	0.20	0.18
5	24.19	24.59	24.63	0.16	0.18	0.16
6	23.37	23.90	24.01	0.14	0.16	0.13
7	22.74	23.24	23.46	0.13	0.14	0.12
8	22.23	22.64	23.01	0.12	0.13	0.11
9	21.88	22.22	22.60	0.11	0.11	0.11
10	21.53	21.73	22.31	0.10	0.11	0.10
15	19.98	20.34	20.75	0.08	0.08	0.08
20	18.84	19.22	19.70	0.06	0.07	0.07
25	17.94	18.26	18.88	0.05	0.06	0.06
30	17.18	17.48	18.14	0.05	0.05	0.05
35	16.51	16.83	17.45	0.04	0.05	0.05
40	15.88	16.21	16.76	0.04	0.04	0.04
45	15.32	15.67	16.15	0.04	0.04	0.04
50	14.84	15.18	15.58	0.03	0.03	0.03
75	12.83	13.06	13.23	0.02	0.02	0.02
100	11.81	11.81	11.81	0.00	0.00	0.00

Table 3--Comparing Average Hit Rates (AHR) and Std Dev (AHR) by Sample Size for Random Samples

	Average Hit Rate (AHR)			Standard Deviation of AHR		
	Sample Size			Sample Size		
Cutoff %	1,600	2,000	2,500	1,600	2,000	2,500
1	26.97	27.18	26.95	0.47	0.49	0.49
2	26.94	26.91	27.45	0.36	0.34	0.33
3	26.45	26.70	27.26	0.28	0.28	0.28
4	25.78	26.08	26.67	0.24	0.24	0.24
5	25.07	25.41	26.04	0.22	0.21	0.21
6	24.42	24.66	25.35	0.19	0.19	0.19
7	23.83	24.10	24.83	0.18	0.17	0.17
8	23.32	23.64	24.23	0.17	0.16	0.16
9	22.89	23.18	23.76	0.16	0.15	0.15
10	22.41	22.79	23.31	0.15	0.14	0.14
15	20.62	20.88	21.29	0.11	0.11	0.11
20	19.27	19.54	19.67	0.09	0.09	0.09
25	18.25	18.48	18.68	0.08	0.08	0.08
30	17.43	17.65	17.79	0.07	0.07	0.07
35	16.72	16.95	17.10	0.06	0.06	0.06
40	16.13	16.33	16.46	0.06	0.06	0.06
45	15.61	15.76	15.90	0.06	0.05	0.05
50	15.12	15.28	15.40	0.05	0.05	0.05
75	13.14	13.28	13.34	0.04	0.04	0.04
100	11.73	11.73	11.73	0.03	0.03	0.03

Table 4--Comparing Average Hit Rates (AHR) and Std Dev (AHR) for Basic Bootstrap, Random Bootstrap, and Random Samples for Sample Size = 2,500

	Average Hit Rate (AHR)			Standard Deviation of AHR		
Cutoff %	Basic Bootstrap	Random Bootstrap	Random Sample	Basic Bootstrap	Random Bootstrap	Random Sample
1	24.94	27.65	26.95	0.40	0.50	0.49
2	25.22	26.97	27.45	0.27	0.35	0.33
3	25.65	26.33	27.26	0.22	0.29	0.28
4	25.33	25.59	26.67	0.18	0.25	0.24
5	24.63	24.94	26.04	0.16	0.22	0.21
6	24.01	24.43	25.35	0.13	0.19	0.19
7	23.46	23.78	24.83	0.12	0.18	0.17
8	23.01	23.25	24.23	0.11	0.16	0.16
9	22.60	22.83	23.76	0.11	0.15	0.15
10	22.31	22.46	23.31	0.10	0.14	0.14
15	20.75	20.74	21.29	0.08	0.11	0.11
20	19.70	19.30	19.67	0.07	0.09	0.09
25	18.88	18.29	18.68	0.06	0.08	0.08
30	18.14	17.44	17.79	0.05	0.07	0.07
35	17.45	16.73	17.10	0.05	0.07	0.06
40	16.76	16.10	16.46	0.04	0.06	0.06
45	16.15	15.56	15.90	0.04	0.06	0.05
50	15.58	15.07	15.40	0.03	0.05	0.05
75	13.23	13.06	13.34	0.02	0.04	0.04
100	11.81	11.73	11.73	0.00	0.03	0.03

Table 5-Comparing Average Hit Rates (AHR) by Sample Size for Basic Bootstraps and Random Samples

	Basic Bootstraps			Random Samples		
	Sample Size			Sample Size		
Cutoff %	1,600	2,000	2,500	1,600	2,000	2,500
1	26.04	24.67	24.94	26.97	27.18	26.95
2	25.36	25.16	25.22	26.94	26.91	27.45
3	25.11	25.48	25.65	26.45	26.70	27.26
4	24.77	25.18	25.33	25.78	26.08	26.67
5	24.19	24.59	24.63	25.07	25.41	26.04
6	23.37	23.90	24.01	24.42	24.66	25.35
7	22.74	23.24	23.46	23.83	24.10	24.83
8	22.23	22.64	23.01	23.32	23.64	24.23
9	21.88	22.22	22.60	22.89	23.18	23.76
10	21.53	21.73	22.31	22.41	22.79	23.31
15	19.98	20.34	20.75	20.62	20.88	21.29
20	18.84	19.22	19.70	19.27	19.54	19.67
25	17.94	18.26	18.88	18.25	18.48	18.68
30	17.18	17.48	18.14	17.43	17.65	17.79
35	16.51	16.83	17.45	16.72	16.95	17.10
40	15.88	16.21	16.76	16.13	16.33	16.46
45	15.32	15.67	16.15	15.61	15.76	15.90
50	14.84	15.18	15.58	15.12	15.28	15.40
75	12.83	13.06	13.23	13.14	13.28	13.34
100	11.81	11.81	11.81	11.73	11.73	11.73

Table 6--Shapiro-Wilk Test of Normality of Basic Bootstraps, Random Bootstraps, and Random Samples for Sample Size = 2,500

	Shapiro-Wilk Test Statistic (W)			Significance Level (Prob < W)		
Cutoff %	Basic Bootstrap	Random Bootstrap	Random Sample	Basic Bootstrap	Random Bootstrap	Random Sample
1	0.9726	0.9834	0.9857	<0.0001	0.0001	0.0005
2	0.9863	0.9776	0.9911	0.0008	<0.0001	0.0163
3	0.9909	0.9910	0.9927	0.0142	0.0158	0.0486
4	0.9906	0.9889	0.9962	0.0117	0.0039	0.4526
5	0.9923	0.9913	0.9956	0.0361	0.0189	0.3306
6	0.9946	0.9937	0.9951	0.1782	0.0934	0.2325
7	0.9953	0.9922	0.9971	0.2686	0.0346	0.7127
8	0.9926	0.9931	0.9972	0.0465	0.0641	0.7279
9	0.9953	0.9923	0.9978	0.2751	0.0360	0.8746
10	0.9964	0.9948	0.9966	0.4992	0.1905	0.5654
15	0.9950	0.9970	0.9945	0.2265	0.6669	0.1595
20	0.9969	0.9979	0.9974	0.6492	0.9127	0.7917
25	0.9971	0.9954	0.9962	0.7115	0.2797	0.4526
30	0.9954	0.9965	0.9966	0.2865	0.5396	0.5735
35	0.9946	0.9932	0.9973	0.1733	0.0699	0.7626
40	0.9970	0.9916	0.9971	0.6777	0.0224	0.6977
45	0.9955	0.9962	0.9973	0.2966	0.4650	0.7658
50	0.9925	0.9968	0.9980	0.0424	0.6235	0.9160
75	0.9926	0.9970	0.9967	0.0443	0.6641	0.5878