Evaluating Modeling Error of Imputed IRS Income Proportions Using Balanced Bootstrap on Simulated Variables

Chih-Chin Ho and William Wong, Internal Revenue Service

In IRS a sample of Tax Year 1988 individual tax returns contains the differences between the examiner-determined value (E) and the taxpayer-reported values (R) for each of 15 income sources. Portions of these differences (D = E - R) are detectable from information documents such as wage and interest statements. These information document portions are available for the 54,088 timely-filed returns but not for the 2,208 delinquent returns.

For this study, interest income portions (y_0) are modeled using logistic regression from the 8,173 timely-filed returns having positive interest differences (D>0). The resulting model is then used to impute the portions to the 121 delinquent returns that had positive differences. Both the imputed microdata portions and their averages are used for economic modeling of tax compliance.

To measure the variance, we use 100 sets of imputations from 100 balanced bootstrap samples. Calculating the mean square error (MSE) is more challenging. Here, for each delinquent filer, we find a nearest neighbor matching timely filer. We estimate the bias by imputing to these nearest neighbors and comparing the imputed values with the true values. Adding the squared bias to the variance yields the mean square error.

To determine the accuracy of the MSE estimates, we create two "similar" variables in which the true values for the delinquent filers are known. We then repeat the imputation and error estimation procedures and compare our MSE estimates with those based on the true values.

■ General Methodology

Upon examining the timely-filed returns for the portions of interest income for which IRS has information documents, we noticed that 7,788 of the 8,173 returns (or 95 percent) had information document portions of either zero or one. Assuming all the information document portions are zero or one allows us to use logistic regression for our imputation.

Regression Model Based Imputation

First, a logistic regression is run on the timely-filed returns to model the information document portion for interest income. The model is then applied to the delinquent filers to impute the portions. Since it is unclear whether it is preferable to use fractional imputed portions or to have them converted to zeros and ones, both cases are studied.

Variance and MSE Estimation

Estimates of the average imputed portion for delinquent filers contain both sampling error and imputation error. Thus, estimates of their variance and mean square error need to contain both sampling and imputation error. Individual microdata imputed portions contain only imputation error. Here we measure only the variance and mean square error due to imputation.

To measure the mean square errors, we measure the squared bias and the variance and add them together. To measure the bias, each delinquent filer is matched to a timely filer. These timely filers, called pseudo-copies, act as surrogates for the delinquent filers. Imputing to these pseudo-copies provides an estimate of the bias. To measure the variance, 100 balanced bootstrap samples are drawn and 100 logistic regression models are computed. These 100 models are then used to create 100 sets of imputations to both the delinquent filers and the pseudo-copies. The variances and MSEs can now be calculated.

Simulated Variables

To determine the accuracy of the mean square error estimation procedure, we create variables that are similar to the variable we tried to impute, y_0, the information portion for interest. Here, however, we create variables that can be calculated for the delinquent filers.

To create our first simulated variable, y_1 , we first divide the taxpayer-reported interest by the examiner-determined interest. Most of these ratios are neither zero nor one, whereas our original variable, y_0, is zero or one 95 percent of the time. (This was the reason we used logistic regression.) To correct this we forced all but the 385 lowest nonzero ratios to one. The fractional values were then ratio adjusted upward.

Our second simulated variable, y_2 , is the same ratio used in y_1 but without the correction.

Since we have the true values here, we can determine the accuracy of our pseudo-copy MSE estimates.

The Imputation Regression Model

Original Variable

To model the information document portions for interest income, $y_0(n)$, for delinquent filer, n, SAS "fast backwards elimination" logistic regressions with a "significance level of staying" of 0.05 were run on the timely-filed returns. The modeling variables, $x_i(n)$, were: the intercept; nine of ten occupation class indicators; nine of ten examination class indicators; the interest D (= E - R); the interest D/ E ratio; the interest E / total income E ratio; the interest D / total income D ratio; the squares of the last four amounts; and for each of the 15 income variables, an indicator variable of whether the income was positive and an indicator variable of whether the income was negative.

Simulated Variables

Modeling the simulated variables, $y_1(n)$ and $y_2(n)$, is carried out in a similar fashion but with different sets of modeling variables.

For $y_1(n)$, the modeling variables were: the intercept; nine of ten occupation class indicators; nine of ten examination class indicators; and for 13 of the 15 income variables, an indicator variable of whether the income was positive and an indicator variable of whether the income was negative.

For $y_2(n)$, the modeling variables were: the intercept; nine of ten occupation class indicators; nine of ten examination class indicators; the ratio of examiner-determined interest divided by the total examiner-determined income; and indicator variables for six income types.

Creating Pseudo-Copies

Pseudo-copies (PCs) are timely filers (TFs) that act as surrogates for the delinquent filers (DFs). They are needed to estimate the mean square errors. For each of the three variables, a set of 121 pseudo-copies are created as follows:

- ❑ Run SAS fast backwards elimination regression on the 8,173 TF returns based on the same independent variables as is used for logistic regression imputation.
- ❑ Apply the resulting model to each of the 8,173 TF returns and the 121 DF returns to obtain a match variable.
- ❑ Find for each DF return a nearest neighbor TF return, using the match variable.

Creating Bootstrap Samples

A set of balanced bootstrap samples are selected for each of the three variables. They are created from sampling the remainder (RTF) of the 8,173 TF returns less the 121 PC returns.

The method used to select balanced bootstrap samples was introduced by Davison, Hinkley, and Schechtman (1986) and described in Hall (1992).

One hundred balanced bootstraps samples are selected from the RTF as follows:

- □ Create a string of B=100 identical copies of the RTF. Thus, the string contains B**n* units where n is the number of returns in RTF.
- □ Randomly permute the units in this string. This can be done by assigning each return a random number and then sorting by it.
- □ The first *n* units are bootstrap 1, the second n units are bootstrap 2, ... , and the last *n* units are bootstrap 100.

■ Creating Bootstrap Imputations

Bootstrap imputations for the DF and PC returns are calculated based on each of the 100 bootstrap samples from the RTF as follows:

- □ Convert the variable (y_0, y_1, or y_2) to zero or one and obtain the logistic regression model coefficients for each bootstrap sample. For y_0 and y_1, all nonzero values are set to one. For y_2, all values greater than 0.5 are set to one.
- □ Calculate the logits by applying these coefficients to the DF returns.
- □ Invert the logits to obtain the fractional imputed values.
- □ For the study of the non-fractional imputed value case, convert the fractional imputed values to ones or zeros depending on whether they are greater than uniform (0,1) random numbers, R(b,n), for bootstrap (b) by return (n).
- □ Apply the last three steps to the PC returns.

■ Variance Estimation

Variances of Individual Imputations

For the individual imputations, the variances calculated include only the imputation variation, not the sampling variation from the population variance.

The variance estimate of the original model individual imputations to the delinquents for the original variable, y_0, is:

$$V(y_0) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{B-1}\sum_{b=1}^{B} (y_0(b,n) - \bar{y}_0(n))^2 ,$$

where b denotes bootstrap, n denotes the nth delinquent file return, and $\bar{y}_0(n)$ is the average across bootstraps.

This estimate applies to both imputing fractions as well as the converted imputations. Also, the variances for the simulated variables, y_1 and y_2, are similarly defined.

Variances of the Average Estimates

Estimates of the average information document portions contain both sampling and imputation error. Their estimates of variance need to include both these errors.

The variance estimate of the original model estimate of the average,

$$\bar{y}_0 = \frac{1}{N}\sum_{n=1}^{N} y_0(n) ,$$

for the original variable, y_0, is:

$$V(\bar{y}_0) = \frac{1}{B-1}\sum_{b=1}^{B} (\bar{y}_0(b) - \bar{y}_{0\cdot})^2 ,$$

where $\bar{y}_0(b)$ is the average over all delinquent filers n for fixed bootstrap $\bar{y}_0$ and y_0 is the average of $\bar{y}_0(b)$ across bootstraps.

Again, this applies to both imputing fractions as well as the converted imputations. Also, the estimators and their variances for the simulated variables, y_1 and y_2, are similarly defined.

These bootstrap variances are slight overestimates, since balanced bootstraps of size n were used instead of independent bootstraps of size n-1. Also, finite population corrections and stratification differences are assumed to be small.

■ Mean Square Error Estimation

The mean square errors (MSEs) of the imputations cannot be directly calculated. However, if imputations to the pseudo-copies are good proxies for imputations to the delinquents, then the mean square errors can be estimated from the pseudo-copies.

MSEs of Individual Imputations

For the individual imputations, the mean square errors can be estimated in two ways.

First, the mean square error can be estimated by applying the original logistic regression model to the pseudo-copy and calculating the mean square difference between the pseudo-copy imputed value and its true value. (E.g.,

$$MSE_1(y_0) = \frac{1}{N}\sum_{n_p=1}^{N} (y_0(n_p) - y_{0T}(n_p))^2 ,$$

where $y_0(n_p)$is the pseudo-copy original model imputed value and
$y_{0T}(n_p)$is the pseudo-copy true value.)

A second estimate can be obtained by making that calculation for each bootstrap and taking the average. (E.g.,

$$MSE_2(y_0) = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{N}\sum_{n_p=1}^{N} (y_0(b,n_p) - y_{0T}(n_p))^2 ,$$

where $y_0(b,n_p)$ is the pseudo-copy bootstrap b model imputed value.)

For the individual imputations for variables y_1 and y_2, the true mean square error is calculated from the delinquent filers. (E.g.,

$$MSE_T(y_1) = \frac{1}{N}\sum_{n_d=1}^{N} (y_1(n_d) - y_{1T}(n_d))^2 ,$$

where n_d denote delinquent file returns.)

MSEs of the Average Estimates

For the estimates of averages, the mean square error can be estimated three ways.

First, after applying the original logistic regression model to the pseudo-copy, the mean square error can be estimated by calculating the variance of the mean for the pseudo-copy and adding the squared bias from the pseudo-copy. (E.g., for

$$\bar{y}_0 = \frac{1}{N}\sum_{n=1}^{N} y_0(n) ,$$

$$MSE_1(y_0) = \left[\frac{1}{N}\frac{1}{N-1}\sum_{n_p=1}^{N} (y_0(n_p) - \bar{y}_{0_p})^2\right] + [\bar{y}_{0_p} - \bar{y}_{0T_p}]^2 ,$$

where $\bar{y}_{0_p}$ is the average of the $y_0(n_p)$ and $\bar{y}_{0T_p}$ is the average of the pseudo-copy true values.)

A second estimate is obtained by making the above calculation for each bootstrap and then averaging the bootstrap estimates. (E.g.,

$$MSE_2(y_0) = \frac{1}{B}\sum_{b=1}^{B} \{[\frac{1}{N}\frac{1}{N-1}\sum_{n_p=1}^{N} (y_0(b,n_p) - \bar{y}_{0\cdot_p}(b))^2] + [\bar{y}_{0_p}(b) - \bar{y}_{0T_p}]^2\} ,$$

where $\bar{y}_{0.,p}(b)$ is the average of the $y_0(b,n_p)$.)

The third estimate is obtained by replacing the variance of the mean in the first estimate by the bootstrap estimate of the variance. (E.g.,

$$MSE_3(y_0) = \frac{1}{B-1}\sum_{b=1}^{B}\{ [(\bar{y}_{0_p}(b) - \bar{y}_{0.,p})^2] + [\bar{y}_{0.,p} - \bar{y}_{0T_p}]^2 \},$$

where $\bar{y}_{0.,p}$ is the average of the $\bar{y}_{0.,p}(b)$.)

For estimates of averages for variables y_1 and y_2 the true mean square error is calculated from the delinquent filers by adding the variance of the mean to the squared bias. (E.g.,

$$MSF_T(y_1) = \left[\frac{1}{N}\frac{1}{N-1}\sum_{n_d=1}^{N}(y_1(n_d) - \bar{y}_{1.d})^2\right] + [\bar{y}_{1_d} - \bar{y}_{1T_d}]^2 \text{ .)}$$

■ Results

The results are given in Table 1. Initially, reviewing only the results for the original variable, y_0 , it appeared that we successfully estimated at least part of the bias and may have reasonable estimates of the root mean square error (RMSE). At that time, we did not know how successful the pseudo-copying procedure was. We proceeded to simulate variables to find out. The analysis below of the simulated variables indicates that our success may, indeed, be limited and that care must be taken in determining which variable to analyze and how to proceed with the matching to create the pseudo-copies.

Mean Values

For the original variable, y_0 , we notice that the pseudo-copy true mean of 0.810 is slightly larger than the timely filer true mean of 0.804. This indicates that pseudo-copying may be picking up some of the characteristics of the delinquent filers. We also notice that the pseudo-copy imputed mean of 0.851 differed from the pseudo-copy true mean. This gave us an estimate of bias of 0.041. The same bias was obtained whether we imputed fractions or converted them to zeros and ones. The question remains as to what proportion of the bias we actually captured.

For the first simulated variable, y_1, we notice that the true mean of the delinquent filers (0.360) was substantially less than the that of the pseudo-copy (0.582), which was, in turn, less than that of the timely filers (0.681). This shows that pseudo-copying has captured some, but not all, of the characteristics of the delinquent filers. The disappointing news is that for imputing fractions, we did not capture any bias (0=0.582-0.582) in the pseudo-copy when a significant true bias (0.221=0.581-0.360) exists. For imputing zero-one's, it appears that a portion (0.030=0.612-0.582) of the bias (0.252=0.612-0.360) has been captured. However, it is more likely that we captured rounding variation rather than bias.

For the second simulated variable, y_2, analysis of the true means again indicates that pseudo-copying has captured some of the characteristics of the delinquent filers. Here, for imputing fractions, we also captured part (0.059=0.494-0.435) of the bias (0.264=0.492-0.228). We speculate on two reasons why we captured bias here but did not for the first simulated variable. First, the ordinary least squares (OLS) regression matching may have picked up the non zero-one nature of some of the observations that was not picked up by the logistic regression modeling. Second, quantitative variables were used to model this simulated variable but not for the first simulated variable.

Standard Deviations

As expected, the standard deviations for imputing fractions are substantially lower than those for imputing zero-one's. The standard deviations are similar across the three variables.

Table 1. Means, SD's, and Root MSE's of Imputing Fractions/Zero-Ones's for Individual and Average Estimates for the Original & 2 Simulated Variables

	Timely Filer	Delinquent Filer		Pseudo Copy				Original Model	Ave. of Bootstrap	Bootstrap
Method	True Mean	True Mean	Imp. Mean	True Mean	Imp. Mean	Estim. Std Dev	True Rt MSE	Estim. Rt MSE	Estim. Rt MSE	Estim. Rt MSE
Analysis of Original Variable Y0 (Interest: Infomation Document Portion):										
Individual Records:										
Imputing Fractions	0.804	n.a.	0.857	0.810	0.851	0.025	n.a.	0.394	0.395	N/A
Imputing Zero-One's	0.804	n.a.	0.884	0.810	0.851	0.343	n.a.	0.538	0.533	N/A
Estimating Averages:										
Imputing Fractions	0.804	n.a.	0.857	0.810	0.851	0.007	n.a.	0.054	0.053	0.041
Imputing Zero-One's	0.804	n.a.	0.884	0.810	0.851	0.030	n.a.	0.042	0.049	0.054
Analysis of Simulated Variable Y1 (Interest: Adjusted Voluntary Reporting Percentage):										
Individual Records:										
Imputing Fractions	0.681	0.360	0.581	0.582	0.582	0.031	0.469	0.441	0.443	N/A
Imputing Zero-One's	0.681	0.360	0.612	0.582	0.612	0.433	0.687	0.640	0.620	N/A
Estimating Averages:										
Imputing Fractions	0.681	0.360	0.581	0.582	0.582	0.009	0.225	0.044	0.045	0.008
Imputing Zero-One's	0.681	0.360	0.612	0.582	0.612	0.036	0.252	0.036	0.046	0.050
Analysis of Simulated Variable Y2 (Interest: Voluntary Reporting Percentage):										
Individual Records:										
Imputing Fractions	0.557	0.228	0.492	0.435	0.494	0.027	0.411	0.375	0.379	N/A
Imputing Zero-One's	0.557	0.228	0.438	0.435	0.512	0.455	0.587	0.616	0.595	N/A
Estimating Averages:										
Imputing Fractions	0.557	0.228	0.492	0.435	0.494	0.007	0.266	0.071	0.073	0.059
Imputing Zero-One's	0.557	0.228	0.438	0.435	0.512	0.038	0.211	0.079	0.078	0.087

Notes: n.a. True values not available for the original variable.
N/A Bootstrap estimates of root MSE not applicable to individual record estimates.

Root Mean Square Errors

For imputing individual records, two methods of estimating the root mean square error (RMSE) were available. Both methods seemed to do equally well. They both had a small downward bias. For imputing fractions for the first simulated variable, the RMSE estimates were around 0.44, whereas the true RMSE was 0.47. Results were similar for the second simulated variable. For imputing zero-one's, the picture is not as clear, due to the added conversion variation. This suggests that the RMSE estimates for the original variable are usable, though they may be slightly biased downward. There is no preference between the two RMSE estimates. The original model estimate may be less biased, but is likely to have more variance.

For estimating averages, three methods of estimating the RMSE were available. The bootstrap estimate, MSE3 , appears to be less stable. For example, for imputing fractions for the first simulated variable, a value of 0.008 is unrealistically low. The other two estimates substantially underestimate the RMSE. For imputing fractions for the first simulated variable, estimated RMSE values around 0.045 were well below the true value of 0.225. For the second simulated variable the estimate of 0.07 is proportionately closer to the real value of 0.266. The cause is the inability of the pseudo-copying to estimate the bias. This was discussed in the mean values section. Again, there is no preference between the two estimates. Thus, for the original variable, there is considerable likelihood that the RMSE estimates are severe underestimates.

■ Conclusions

It appears that creating simulated variables and evaluating them was a very valuable experience. It showed that the root mean square error estimates for the individual estimates are likely to be usable, but not for estimating averages. It showed that the pseudo-copying technique had weaknesses. Care must be taken in determining which variables to analyze using pseudo-copying, what method to use for matching, and which variables to use in the regressions. This study showed just how difficult estimating bias can be. It reemphasizes that only obtaining the true values will tell us how much bias remains.

■ Recent Developments

Prompted by the poor estimates of bias and mean square error for the first simulated variable, we sought to improve the pseudo-copy matching. We tried a more elaborate set of independent variables for our OLS regression matching. The results showed some promise. We were now able to capture part (0.06=0.582-0.522) of the bias (0.221=0.581-0.360). Consequently, our measures of mean square error also improved. For imputing fractions for the first simulated variable, estimated RMSE's were now around 0.077 (up from 0.008), which is closer to the real value of 0.225.

■ Future Research

To complete this study, ideally, we would try to obtain the real information document portions for the delinquent filers. This is not possible. The study of the simulated variables suggests that alternative methods of matching with perhaps different sets of independent variables be studied. It also suggests that we try to determine some criteria to evaluate when we expect the method will succeed and what percent of the bias we may anticipate obtaining.

■ Acknowledgments

The authors would like to thank Dennis Cox for his review and suggestions.

■ References

Davison, A.C.; Hinkley, D.V.; and Schechtman, E. (1986). Efficient Bootstrap Simulation, *Biometrika* 73, 555-566.

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

Selected Bibliography

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

Ho, Chih-Chin and Wong, William (1994). Alternative Imputation Techniquès for Proportions of Income Variables for IRS Compliance Modeling, in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1, 388-393.

Ho, Chih-Chin and Wong, William (1995). Measuring Modeling Error and Variances of Imputed IRS Income Proportions Using Balanced Bootstrap and Multiple Imputation, *1995 Proceedings of the International Conference on Survey Measurement and Process Quality*, American Statistical Association, 314-319.

Shao, Jun and Tu, Dongsheng (1995). *The Jackknife and Bootstrap*, Springer-Verlag. ■