ALTERNATIVE IMPUTATION TECHNIQUES FOR PROPORTIONS OF INCOME VARIABLES FOR IRS COMPLIANCE MODELING

Chih-Chin Ho, William Wong, Internal Revenue Service
Chih-Chin Ho, IRS, 1111 Constitution Ave., Washington, D.C. 20224

KEY WORDS: Nearest Neighbor Hot Deck, Regression, Cross Validation

In IRS a sample of individual income tax returns is subject to a detailed line-by-line audit by IRS Examination. For each of 15 income sources the difference between the examined value and the taxpayer-reported value is calculated. A portion of this difference is detectable from information reports, such as wage and interest statements. These portions are used in economic models of tax compliance. For a file of delinquent returns the portion of the difference detectable through information reports was not available. We sought methods to impute estimates of these portions from timely filed data.

Several primary methods of imputation are considered: regression, nearest neighbor hot deck imputation, and imputing cell means. Various approaches to these methods using different stratifications and different variables are tried. Since the true portions for the delinquent returns were not available for any of the returns, indirect methods of evaluation were needed. This paper compares the methods using half sample cross-validation.

BACKGROUND

The timely filer file consisted of a sample of 54,088 Tax Year 1988 returns. For each return both the taxpayer-reported amount (Y1) and the examined amount (Y2) were available for each of 15 income types. The detected amount (Y4=Y2-Y1) is then the difference between the examined amount and the taxpayer-reported amount for each type of income. Also available here was the portion of the detected income that IRS attributed to information documents such as wage statements (Forms W-2) and interest statements (Forms 1099). Each return also had the auxiliary variables occupation and examination class.

The delinquent filer file consisted of a sample of 2,208 Tax Year 1988 returns. Again, both the taxpayer-reported amount and the examined amount were available for each of 15 income types and the detected amount could then be calculated as Y4=Y2-Y1. Here, however, the portion of the detected income attributed to information documents was not available and needed to be imputed. For simplicity, we considered only cases where the portions were between zero and one, inclusive.

METHODOLOGY

For each income variable, the timely filer file was split into two halves by alternately assigning returns to half samples HA and HB, after removing returns that had zero detected income, since calculating information document portions of zero makes little sense. The procedure was then to use one of the half samples, say HB, to do the modeling, such as calculating cell means, and then apply the resulting information document portions to the other half sample, HA. Since the true value of the portion also resides on the other half sample, the absolute and square differences between the true and imputed values provide measures of the effectiveness of the procedures. Now, by reversing the roles of HA and HB, a second set of evaluations can be calculated. Comparing the pairs of evaluations yields a rough measure of the stability of the imputation procedures. This methodology is then applied to the three main imputation procedures on each of three income variables. The first income variable selected (interest) had a moderately high information document portion, the second income variable (other income/loss) had a moderately low portion, and the third income variable (Schedule E income/loss) had a very low portion.

IMPUTATION PROCEDURES

A. Cell Mean Imputation

For the first mean procedure (M1), we start by calculating the overall mean information document portion across the entire half sample. This one mean is then imputed to every return in the second half sample.

For the second mean procedure (M2), we calculate separate mean information document portions for each of 10 examination classes and then impute them to the corresponding examination class in the second half sample. Examination classes basically consist of the form type by total positive income or total receipts.

For the third mean procedure (M3), we calculate separate mean information document portions for each of 10 occupation classes and then impute them to the corresponding occupation class in the second half sample.

Preliminary work showed that using more detailed examination or occupation classes resulted in higher mean square errors.

B. Nearest Neighbor Hot Deck

For the first nearest neighbor hot deck procedure (N1), we sort both half samples HA and HB by the taxpayer-reported amount (Y1). In using HB to impute into HA, for each return in HA, we find the record in HB whose taxpayer-reported amount (X1) is closest to Y1 and impute X1's information document portion to the HA record. When there are multiple exact matches, we select a systematic sample.

For the second nearest neighbor hot deck procedure (N2), we sort both half samples HA and HB by the taxpayer-reported amount (Y1) within examination class. In using HB to impute into HA, for each return in HA, we find the record in HB in the same examination class whose taxpayer-reported amount (X1) is closest to Y1 and impute his information document portion to the HA record. Again, we systematically sample multiple exact matches.

For the third nearest neighbor hot deck procedure (N3), we repeat (N2) replacing examination class with occupation class.

For the fourth, fifth, and sixth procedures (N4, N5, and N6), we repeat the first three procedures use the examined amounts (Y2 and X2) instead of the taxpayer-reported amounts (Y1 and X1).

For the seventh, eighth, and ninth procedures (N7, N8, and N9), we repeat the first three procedures using the detected amounts (Y4=Y2-Y1 and X4=X2-X1) instead of the taxpayer-reported amounts (Y1 and X1).

For the tenth procedure (N10), when imputing from HB to HA, we calculate a logistic regression model from HB and apply the model to both HA and HB, to obtain *logit* values for each record in HA and HB. We now use the *logits*, instead of the taxpayer-reported amounts, to perform a nearest neighbor hot deck.

C. Regression

For the full model regression procedure (R1), we calculate a logistic regression from one half sample, HB, and apply the model to the other half sample, HA. logistic regression was repeated on a variety of modeling variables until a basic set of significant variables was obtained. For modeling the portion for the first income variable, interest, the final modeling variables were: the intercept; nine occupation class indicators; nine exam class indicators; the interest Y4 difference; the interest ratio Y4/Y2; the ratio of the interest Y2 / total income Y2; the ratio of the interest Y4 / total income Y4; the squares of each of the four interest income terms above; and, for all of the income variables, indicator variables of whether the income was positive and whether it was negative. A detailed investigation helped explain why, whenever a variable was significant, so was its quadratic term. Similar models were used to model the information document portions for the other two income variables. To perform the regression, the dependent variable (the information document portion for the income variable) was set to one whenever it was greater than zero. Typically, only a small fraction of the returns had portions not zero or one. (When calculating the evaluation statistics, the imputed portion was compared to the true portion, instead of the adjusted portion.)

For the short model regression procedure (R2), we applied a backwards elimination procedure to the full model in (R1), using a significance level of 0.2 to yield around 10 modeling variables.

For the redistributed full and short model regression procedures (R3 and R4), we tried to modify the regressions to reflect the distributions of the portions in the modeling half sample. For R3, after calculating the R1 model from HB and applying it to HA, HA was then sorted by the logit value, and HB was sorted by the information document portion and the distribution of portions in HB were translated over to HA. For the procedure R4, the R2 model was used instead of R1.

EVALUATION CRITERIA

To evaluate the different imputation procedures, three criteria were used:

1. Absolute Bias =

$$\frac{|\Sigma \textit{Imputed Portion} - \Sigma \textit{True Portion}|}{\textit{Number of Observations}}$$

2. Mean Absolute Error =

$$\frac{\Sigma |\textit{Imputed Portion} - \textit{True Portion}|}{\textit{Number of Observations}}$$

3. Mean Square Error =

$$\frac{\Sigma(\textit{Imputed Portion} - \textit{True Portion})^2}{\textit{Number of Observations}}.$$

For each imputation method two half sample estimates were computed to give us an indication of the variability of the methods.

RESULTS

The imputation was tested on three income variables: Interest Income (V1), Other Income/Loss (V2), and Schedule E Income/Loss (V3). The results are illustrated in Figures 1, 2, and 3, and presented in Tables 1, 2, and 3, respectively.

A. Interest Income

For Interest Income, the full model and short model regressions (R1 and R2) had lower mean square errors than all the other procedures. Both procedures had mean square errors of around 0.14, whereas the mean imputation procedures had slightly higher mean square errors of around 0.15. The mean square errors of all the nearest neighbor procedures and the redistributed regression procedures were twice as high. Since almost all of the original portions were zero or one, the doubling of the mean square error for the nearest neighbor procedures should have been expected. A theoretical explanation of this factor of two is given in the Appendix. The regression procedures had a smaller mean absolute error than the mean procedures, but had a larger absolute bias. For this variable, regression imputation is recommended. The short regression model is preferred since it is easier to explain economically.

B. Other Income/Loss

For Other Income/Loss, the results are very similar to Interest Income. Both regression models had mean square errors of around 0.16, beating the mean procedures mean square errors of 0.17. Here, however, the regression procedures also beat the mean procedures in both lower mean absolute errors and absolute bias. For this variable, regression imputation is the clear favorite.

C. Schedule E Income/Loss

For Schedule E Income/Loss, the mean imputation procedures edged out the regression procedures in mean square error, mean absolute error and absolute bias. Only around 100 of the returns in each half sample had non-zero portions. This made the regression models rather unstable. The stability of mean imputation proved to be more important than the potential gain from using many variables in the regression. Thus, mean imputation is recommended here.

CONCLUSIONS

The full and short model regression procedures appear to yield the smallest mean square error, except when there is insufficient data to stabilize the model. When either the number of observations with portions of zero or one is less than 100, the stability of the model may be suspect. In such cases mean imputation is preferred. With almost all the data having portions of zero or one, the redistributed regression and all the nearest neighbor procedures are never preferred, since their mean square errors will be twice as high as the regression or mean procedures. This fact is demonstrated in the Appendix.

FUTURE RESEARCH

Instead of splitting the timely filer sample into two half samples, we can more closely mimic the size and variable by variable distribution of the delinquent taxpayer file using a subsample of one half sample and study how well imputation procedures or models derived from the other half sample work. Variables can be studied individually or collectively.

We plan to continue this imputation investigation on the remaining 12 variables. This would give us an indication of which methods are consistently superior and under which conditions. It will also give us further indications of the variability of our procedures and results.

The best and perhaps the only valid evaluation is to obtain the true information document portions from the actual records of delinquent filers for which we are trying to impute. Failing this, one alternative is to repeat this procedure on another year of data, where the information document portions are available for the delinquent returns.

ACKNOWLEDGMENTS

The authors would like to thank Dennis Cox for his review and suggestions and Wendy Alvey for her assistance in preparing the paper and its presentation.

BIBLIOGRAPHY

1. Kaufman, L. and Rousseeuw, P. J., (1990), *Finding Groups in Data*, John Wiley & Sons, Inc.
2. Little, R. J. A. and Rubin, D. B., (1987), *Statistical Analysis With Missing Data*, John Wiley & Sons, Inc.
3. Rubin, D. B., (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc.

Figure1: V1 (Interest) - A HIGH Information Document Portion Variable

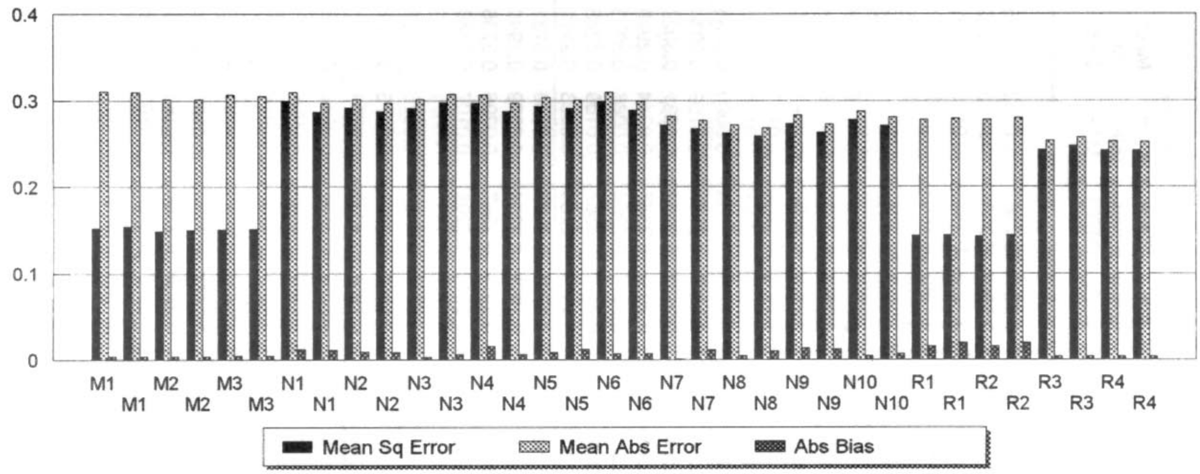

Figure 2: V2 (Other Income/Loss) - A LOW Information Document Portion Variable

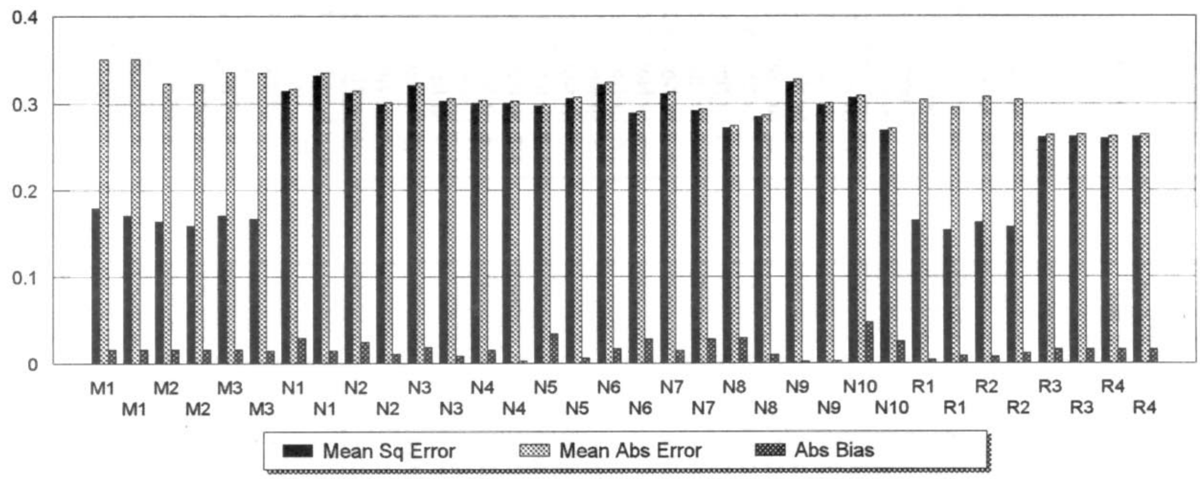

Figure 3: V3 (Schedule E Income/Loss) - A VERY LOW Information Document Portion Variable

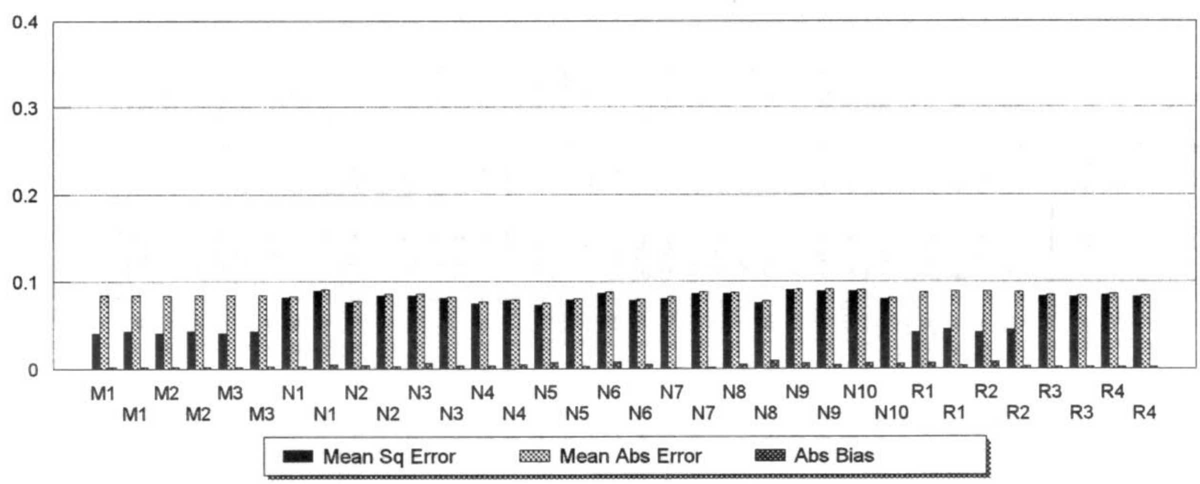

Table 1 - V1 (Interest) - A HIGH Information Document Portion Variable

Imputation Method	Sort	Imp Cell	Half Samp	Mean Imp Portion (MIP)	Mean True Portion (MTP)	Abs Bias = Abs (MIP - MTP)	Mean Abs Error	Mean Sqr Error
M1: Mean	none	none	HA	0.8020	0.8056	0.0036	0.3106	0.1518
M1: Mean	none	none	HB	0.8056	0.8020	0.0036	0.3097	0.1536
M2: Mean	none	Exam	HA	0.8019	0.8056	0.0037	0.3019	0.1482
M2: Mean	none	Exam	HB	0.8056	0.8020	0.0037	0.3013	0.1493
M3: Mean	none	Occ	HA	0.8017	0.8056	0.0039	0.3068	0.1503
M3: Mean	none	Occ	HB	0.8063	0.8020	0.0043	0.3052	0.1513
N1: Nr Nbr	Y1: Txpyr	none	HA	0.7940	0.8056	0.0116	0.3094	0.2999
N1: Nr Nbr	Y1: Txpyr	none	HB	0.8128	0.8020	0.0108	0.2971	0.2871
N2: Nr Nbr	Y1: Txpyr	Exam	HA	0.7971	0.8056	0.0084	0.3016	0.2918
N2: Nr Nbr	Y1: Txpyr	Exam	HB	0.8098	0.8020	0.0078	0.2972	0.2877
N3: Nr Nbr	Y1: Txpyr	Occ	HA	0.8031	0.8056	0.0025	0.3014	0.2910
N3: Nr Nbr	Y1: Txpyr	Occ	HB	0.8076	0.8020	0.0056	0.3074	0.2980
N4: Nr Nbr	Y2: Exam	none	HA	0.7904	0.8056	0.0152	0.3066	0.2973
N4: Nr Nbr	Y2: Exam	none	HB	0.8078	0.8020	0.0059	0.2970	0.2878
N5: Nr Nbr	Y2: Exam	Exam	HA	0.7979	0.8056	0.0077	0.3039	0.2939
N5: Nr Nbr	Y2: Exam	Exam	HB	0.7901	0.8020	0.0119	0.3012	0.2914
N6: Nr Nbr	Y2: Exam	Occ	HA	0.7994	0.8056	0.0062	0.3099	0.2993
N6: Nr Nbr	Y2: Exam	Occ	HB	0.8084	0.8020	0.0064	0.2992	0.2889
N7: Nr Nbr	Y4=Y2-Y1	none	HA	0.8062	0.8056	0.0006	0.2814	0.2722
N7: Nr Nbr	Y4=Y2-Y1	none	HB	0.8126	0.8020	0.0107	0.2773	0.2676
N8: Nr Nbr	Y4=Y2-Y1	Exam	HA	0.8017	0.8056	0.0039	0.2725	0.2627
N8: Nr Nbr	Y4=Y2-Y1	Exam	HB	0.8114	0.8020	0.0094	0.2688	0.2590
N9: Nr Nbr	Y4=Y2-Y1	Occ	HA	0.7926	0.8056	0.0130	0.2832	0.2737
N9: Nr Nbr	Y4=Y2-Y1	Occ	HB	0.8138	0.8020	0.0119	0.2732	0.2633
N10:NrNbr	Regr Logit	none	HA	0.8095	0.8056	0.0039	0.2878	0.2782
N10:NrNbr	Regr Logit	none	HB	0.8088	0.8020	0.0068	0.2812	0.2712
R1: F Regr	none	none	HA	0.8210	0.8056	0.0154	0.2779	0.1427
R1: F Regr	none	none	HB	0.8219	0.8020	0.0199	0.2794	0.1441
R2: S Regr	none	none	HA	0.8206	0.8056	0.0151	0.2782	0.1424
R2: S Regr	none	none	HB	0.8216	0.8020	0.0197	0.2800	0.1441
R3:RF Reg	none	none	HA	0.8020	0.8056	0.0036	0.2536	0.2437
R3:RF Reg	none	none	HB	0.8055	0.8020	0.0036	0.2574	0.2477
R4:RS Reg	none	none	HA	0.8020	0.8056	0.0036	0.2529	0.2430
R4:RS Reg	none	none	HB	0.8055	0.8020	0.0036	0.2524	0.2426

Note: Half sample HA had 4087 observations and HB had 4086 observations.

Table 2 - V2 (Other Income/Loss) - A LOW Information Document Portion Variable

Imputation Method	Sort	Imp Cell	Half Samp	Mean Imp Portion (MIP)	Mean True Portion (MTP)	Abs Bias = Abs (MIP - MTP)	Mean Abs Error	Mean Sqr Error
M1: Mean	none	none	HA	0.2204	0.2367	0.0163	0.3508	0.1796
M1: Mean	none	none	HB	0.2367	0.2204	0.0163	0.3512	0.1707
M2: Mean	none	Exam	HA	0.2205	0.2367	0.0162	0.3231	0.1642
M2: Mean	none	Exam	HB	0.2367	0.2204	0.0162	0.3227	0.1589
M3: Mean	none	Occ	HA	0.2202	0.2367	0.0165	0.3355	0.1707
M3: Mean	none	Occ	HB	0.2349	0.2204	0.0144	0.3352	0.1668
N1: Nr Nbr	Y1: Txpyr	none	HA	0.2071	0.2367	0.0295	0.3167	0.3143
N1: Nr Nbr	Y1: Txpyr	none	HB	0.2352	0.2204	0.0148	0.3348	0.3323
N2: Nr Nbr	Y1: Txpyr	Exam	HA	0.2126	0.2367	0.0241	0.3145	0.3118
N2: Nr Nbr	Y1: Txpyr	Exam	HB	0.2315	0.2204	0.0111	0.3011	0.2988
N3: Nr Nbr	Y1: Txpyr	Occ	HA	0.2179	0.2367	0.0188	0.3238	0.3213
N3: Nr Nbr	Y1: Txpyr	Occ	HB	0.2294	0.2204	0.0090	0.3053	0.3027
N4: Nr Nbr	Y2: Exam	none	HA	0.2208	0.2367	0.0159	0.3034	0.3001
N4: Nr Nbr	Y2: Exam	none	HB	0.2173	0.2204	0.0031	0.3027	0.3005
N5: Nr Nbr	Y2: Exam	Exam	HA	0.2023	0.2367	0.0344	0.2999	0.2978
N5: Nr Nbr	Y2: Exam	Exam	HB	0.2271	0.2204	0.0066	0.3073	0.3053
N6: Nr Nbr	Y2: Exam	Occ	HA	0.2198	0.2367	0.0169	0.3245	0.3216
N6: Nr Nbr	Y2: Exam	Occ	HB	0.1921	0.2204	0.0283	0.2911	0.2887
N7: Nr Nbr	Y4=Y2-Y1	none	HA	0.2221	0.2367	0.0146	0.3131	0.3104
N7: Nr Nbr	Y4=Y2-Y1	none	HB	0.1925	0.2204	0.0279	0.2940	0.2914
N8: Nr Nbr	Y4=Y2-Y1	Exam	HA	0.2071	0.2367	0.0296	0.2736	0.2717
N8: Nr Nbr	Y4=Y2-Y1	Exam	HB	0.2311	0.2204	0.0107	0.2868	0.2843
N9: Nr Nbr	Y4=Y2-Y1	Occ	HA	0.2346	0.2367	0.0021	0.3276	0.3246
N9: Nr Nbr	Y4=Y2-Y1	Occ	HB	0.2176	0.2204	0.0028	0.3006	0.2982
N10:NrNbr	Regr Logit	none	HA	0.1892	0.2367	0.0475	0.3087	0.3059
N10:NrNbr	Regr Logit	none	HB	0.1951	0.2204	0.0254	0.2707	0.2687
R1: F Regr	none	none	HA	0.2319	0.2367	0.0048	0.3038	0.1643
R1: F Regr	none	none	HB	0.2296	0.2204	0.0092	0.2951	0.1538
R2: S Regr	none	none	HA	0.2288	0.2367	0.0079	0.3068	0.1626
R2: S Regr	none	none	HB	0.2327	0.2204	0.0123	0.3039	0.1576
R3:RF Reg	none	none	HA	0.2204	0.2367	0.0163	0.2636	0.2609
R3:RF Reg	none	none	HB	0.2367	0.2204	0.0163	0.2644	0.2617
R4:RS Reg	none	none	HA	0.2204	0.2367	0.0163	0.2623	0.2596
R4:RS Reg	none	none	HB	0.2367	0.2204	0.0163	0.2644	0.2618

Note: Half samples HA and HB both had 849 observations.

Table 3 - V3 (Schedule E Income/Loss) - A VERY LOW Info. Doc. Portion Variable

Imputation Method	Sort	Imp Cell	Half Samp	Mean Imp Portion (MIP)	Mean True Portion (MTP)	Abs Bias = Abs (MIP - MTP)	Mean Abs Error	Mean Sqr Error
M1: Mean	none	none	HA	0.0456	0.0435	0.0021	0.0848	0.0406
M1: Mean	none	none	HB	0.0435	0.0456	0.0021	0.0848	0.0429
M2: Mean	none	Exam	HA	0.0454	0.0435	0.0019	0.0843	0.0404
M2: Mean	none	Exam	HB	0.0438	0.0456	0.0018	0.0847	0.0431
M3: Mean	none	Occ	HA	0.0455	0.0435	0.0021	0.0848	0.0408
M3: Mean	none	Occ	HB	0.0433	0.0456	0.0023	0.0847	0.0431
N1: Nr Nbr	Y1: Txpyr	none	HA	0.0463	0.0435	0.0029	0.0837	0.0822
N1: Nr Nbr	Y1: Txpyr	none	HB	0.0504	0.0456	0.0048	0.0916	0.0900
N2: Nr Nbr	Y1: Txpyr	Exam	HA	0.0392	0.0435	0.0043	0.0777	0.0761
N2: Nr Nbr	Y1: Txpyr	Exam	HB	0.0479	0.0456	0.0024	0.0866	0.0843
N3: Nr Nbr	Y1: Txpyr	Occ	HA	0.0496	0.0435	0.0062	0.0861	0.0844
N3: Nr Nbr	Y1: Txpyr	Occ	HB	0.0419	0.0456	0.0036	0.0823	0.0810
N4: Nr Nbr	Y2: Exam	none	HA	0.0403	0.0435	0.0031	0.0767	0.0747
N4: Nr Nbr	Y2: Exam	none	HB	0.0405	0.0456	0.0051	0.0792	0.0780
N5: Nr Nbr	Y2: Exam	Exam	HA	0.0362	0.0435	0.0073	0.0751	0.0735
N5: Nr Nbr	Y2: Exam	Exam	HB	0.0427	0.0456	0.0029	0.0806	0.0793
N6: Nr Nbr	Y2: Exam	Occ	HA	0.0511	0.0435	0.0076	0.0884	0.0871
N6: Nr Nbr	Y2: Exam	Occ	HB	0.0404	0.0456	0.0051	0.0800	0.078
N7: Nr Nbr	Y4=Y2-Y1	none	HA	0.0441	0.0435	0.0006	0.0825	0.0807
N7: Nr Nbr	Y4=Y2-Y1	none	HB	0.0470	0.0456	0.0014	0.0883	0.0863
N8: Nr Nbr	Y4=Y2-Y1	Exam	HA	0.0486	0.0435	0.0051	0.0879	0.0862
N8: Nr Nbr	Y4=Y2-Y1	Exam	HB	0.0359	0.0456	0.0096	0.0772	0.0755
N9: Nr Nbr	Y4=Y2-Y1	Occ	HA	0.0501	0.0435	0.0066	0.0915	0.0905
N9: Nr Nbr	Y4=Y2-Y1	Occ	HB	0.0498	0.0456	0.0042	0.0915	0.0895
N10:NrNbr	Regr Logit	none	HA	0.0500	0.0435	0.0066	0.0907	0.0891
N10:NrNbr	Regr Logit	none	HB	0.0402	0.0456	0.0054	0.0809	0.0794
R1: F Regr	none	none	HA	0.0497	0.0435	0.0062	0.0880	0.0416
R1: F Regr	none	none	HB	0.0486	0.0456	0.0030	0.0891	0.0454
R2: S Regr	none	none	HA	0.0510	0.0435	0.0076	0.0896	0.0414
R2: S Regr	none	none	HB	0.0483	0.0456	0.0027	0.0886	0.0443
R3:RF Reg	none	none	HA	0.0456	0.0435	0.0021	0.0848	0.0832
R3:RF Reg	none	none	HB	0.0435	0.0456	0.0021	0.0842	0.0826
R4:RS Reg	none	none	HA	0.0456	0.0435	0.0021	0.0862	0.0846
R4:RS Reg	none	none	HB	0.0435	0:0456	0.0021	0.0845	0.0829

Notes:

1. Half samples HA and HB both had 2630 observations.
2. Schedule E (Supplemental Income and Loss) includes Rental Real Estate, Royalties, Partnerships, S Corporations, Estates, Trusts, and Real Estate Mortgage Investment Conduits.

APPENDIX

The following discussion demonstrates why the mean imputation procedure is superior to the nearest neighbor procedure for our data and why you should expect the mean square error of mean imputation to be one half of that of nearest neighbor imputation. (Consequently, since regression has characteristics similar to mean imputation, its mean square error should be similar to that of mean imputation.)

Main Assumption:

Since most of our data have information document portions of zero or one and very few have fractions, assume none of the data have fractions.

A. Simplified Case:

Assume a uniform population n with pn units having portions of 1 and $(1-p)n$ units having portions of 0.

Assume the nearest neighbor procedure assigns pn ones and $(1-p)n$ zeros at random to the population. Then the nearest neighbor total square error

$$\begin{aligned} TSE_{NN} &= p^2 n(1-1)^2 + (1-p)^2 n(0-0)^2 + p(1-p)n(1-0)^2 + (1-p)p\,n(0-1)^2 \\ &= 2p(1-p)n. \end{aligned}$$

is:

But the mean imputation total square error is:

$$\begin{aligned} TSE_{Mean} &= p\,n(1-p)^2 + (1-p)n(0-p)^2 = (1-p)n[p(1-p) + p^2] = (1-p)n\,p \\ &= \frac{TSE_{NN}}{2}. \end{aligned}$$

B. General Case:

Split the population into K homogeneous cells, each having uniform portions p_i $(i = 1, ..., K)$. Now apply the simplified case to each cell and sum across cells. Finally, choosing a large enough K would be a proxy for the population.