Hot-Deck Imputation Procedures for the California Personal Income Tax Sample

Darryl Yamashita, California Franchise Tax Board

The Franchise Tax Board (FTB) administers the Personal Income Tax Laws for the state of California. For tax year 1992, FTB collected $16.6 billion in state personal income tax from over 13.3 million taxpayers. The Research Bureau of FTB uses a two-stage stratified, systematic sampling scheme to select approximately 90,000 taxpayers for its Personal Income Tax (PIT) sample. The PIT Sample data are used to estimate the aggregate amounts of various income fields from both the state and Federal tax formsand to evaluate proposed changes in legislation via the Personal Income Tax Model [1].

Identifying item nonresponse in the PIT Sample can be problematic, since all missing data are recorded as zeros, with zeros also being a legitimate value for many items. Since information from a supporting schedule is recorded on the main form, it is possible to determine if a supporting schedule is missing. Rather than focusing on item nonresponse, entire forms were imputed when they were identified as missing. Hot-deck imputation procedures were developed for the California Schedule CA, California Schedule P, Federal Schedule A, and the Federal Schedule E as the first step in an attempt to increase the accuracy of the PIT Sample.

This paper will present the imputation methodology utilized, discuss its effect on the 1992 PIT Sample, and present further research areas for investigation.

■ Pit Sample

For Tax Year 1992, the PIT Sample consisted of 87,219 tax returns, of which 73,603 were California resident returns. There are over 450 variables from both the state and Federal tax returns recorded for the PIT Sample. Sampling weights are used for aggregate income estimates and the PIT Model estimates.

The default value for all variables in the PIT Sample is zero. Thus, it is problematic to try to determine, variable by variable, if a variable has the value of zero or is missing. However, it is possible to mathematically verify certain line items by using the inherent structure of the forms. The math verification portion of the PIT Model creates residuals -- the difference between the computed amount and the recorded amount -- for each line item which can be verified. These residuals are used in the tax model to adjust the recorded values, so that the model has mathematically accurate data while retaining the actual recorded values.

Rather than focusing on item nonresponse, the imputation procedures developed imputes values for entire missing supporting forms. Checking for missing supporting forms can be easily accomplished by matching fields on either the California Form 540 or the Federal Form 1040 to their corresponding supporting schedules. For example, if the taxpayer itemized his/her deduction, there should be a supporting Form 1040 Schedule A.

The number of missing supporting forms is small relative to the numbers of supporting forms filed. Table 1 lists the number of missing forms and the total number of forms in the PIT Sample. Even though the amount of missing data is small, the effect the missing data have on the estimates can be noticeable, since the line items on supporting forms can often be zero, and the estimates are based on the sample weights.

■ Methodology

Missing values on a tax form lead to an underestimation of the aggregate amounts and, since many supporting forms have legitimate zero values, can lead to incorrect inferences produced by models which rely on accurate data.

TABLE 1: Number of Missing Supporting Schedules for the 1992 Sample

Supporting Schedule	Total #	# Missing	% Missing
State Schedule CA Sub.	46,257	411	0.89
State Schedule CA Add.	22,729	84	0.37
State Schedule CA Ded.	53,621	1,163	2.17
Federal Schedule A	59,288	416	0.70
Federal Schedule E	35,737	217	0.61

There are three well-known methods to analyze data with missing values. Weighting techniques are usually employed when the missing data are unit nonresponse. Imputation procedures are used when missing data are item nonresponse. Model-based techniques can be developed for either unit or item nonresponse.

Imputation procedures are more suited for the particular task at hand, since weight-based techniques are more suited to unit nonresponse and parameter estimation, and model-based techniques can be computationally prohibitive and more difficult to implement into the tax models. Also, using imputation does not require a modification to the current tax model.The notion of imputation is relatively simple. Replace each missing value with a plausible value. Many different procedures have been developed to find "plausible" values for different sampling schemes [2]. Three common single imputation procedures are mean imputation, hot-deck imputation, and regression imputation. Mean imputation replaces the missing values with the mean of the observed values. Hot-deck imputation replaces the missing values with observed values. Regression imputation replaces the missing values with the predicted values, using regression estimates based on the observed values.

The hot-deck imputation procedures developed impute observations from taxpayers with similar attributes and adjust the amounts so that the resulting total of the supporting form matches the amount on the main form.

Hot-deck imputation has three main advantages over other single imputation procedures for this situation. First, since the entire form is imputed, hot-deck imputation preserves the covariance structure of the supporting form by using the structure of the responses of the matching taxpayers. Second, other imputation procedures, such as mean and regression imputation, generally will impute nonzero values into all of the fields of the supporting form. Since some of the fields on a supporting form would be zero (for example, not all filers declare moving expenses on their Federal Schedule A), the aggregate number of taxpayers for all of the fields would be overestimated. Two-stage procedures with a zero-value stage were developed using both mean and regression imputation, but hot-deck imputation proved much easier. (See the further research section for more details.) Finally, hot-deck imputation is easily implemented using the SAS statistical software.

A combination of two different hot-deck methods – the adjustment cell method and the nearest neighbor method – were implemented. Adjustment cell hot-deck imputation allows any missing value to be imputed with any observed value in the same adjustment cell. Nearest neighbor hot-deck imputation defines a metric based on a set of covariates and imputes a missing value with the observed value which minimizes the metric. The imputation procedure constructed defined adjustment cells and used the nearest neighbor within the adjustment cell.

Separate hot-deck imputation procedures were developed for the Form 540 Schedule CA, Form 1040 Schedule A, and Form 1040 Schedule E, which differ slightly to take advantage of the mathematical structure of each schedule. Because of the similarities of the imputation procedures, only the im-

putation procedure developed for the Form 1040 Schedule A will be presented below.

■ 1040 Schedule A Imputation

Procedure

For notational purposes, let $A1_{ij}$, $A2_{ij}$,..., $A26_{ij}$ be lines 1 through 26 on the Schedule A for taxpayer i in adjustment cell j. FTB records lines 1, 4, 5, 6, 7, 9a, 9b, 10, 11, 13–21, 24, 25, and 26. The algorithm for the hot-deck imputation procedure for the Form 1040 Schedule A is:

1. Delete all taxpayers who took the standard deduction. Create adjustment cells using the series code, adjusted gross income (AGI) classification, and the number of dependents claimed. Within each adjustment cell, sort the data by their federal itemized deduction. The sort helped us find the nearest neighbor.

   Let ID_{ij} = The amount of the itemized deduction from the Form 1040 or Form 1040A for taxpayer i in adjustment cell j.

For each taxpayer i, in adjustment cell j, do Steps 2 through 8.

2. Check to see if $A1_{ij}$, $A2_{ij}$,..., $A26_{ij}$ are all zero. If some of them are non zero, then go to the next taxpayer in the adjustment cell. Otherwise continue.

3. Let $A26_{ij} = ID_{ij}$ and mathematically determine the value of itemized deduction before limitations. Call this variable $TOTAL_{ij}$.

4. Find taxpayer $_{i*j}$ in the adjustment cell whose Schedule A exists and itemized deduction is closest to taxpayer$_{ij}$.

5. Impute $A4_{ij}$–$A11_{ij}$, $A16_{ij}$, $A17_{ij}$, $A18_{ij}$, $A24_{ij}$, $A25_{ij}$ with

$$A(K)_{ij} = \frac{A(K)_{i*j}TOTAL_{ij}}{TOTAL_{i*j}},$$

6. Impute $A1_{ij}$ with

$$A1_{ij} = \frac{A1_{i*j}A4_{ij}}{A4_{i*j}}.$$

7. Impute $A13_{ij}$, $A14_{ij}$, and $A15_{ij}$ with

$$A(K)_{ij} = \frac{A(K)_{i*j}A16_{ij}}{A16_{i*j}},$$

   where (K) = 13, 14, 15.

8. Let $A21_{ij} = A24_{ij} + 0.02$ (Federal Adjusted Gross Income)
   Impute $A19_{ij}$ and $A20_{ij}$ with

$$A19_{ij} = \frac{A19_{i*j}A21_{ij}}{A21_{i*j}} \text{ and}$$

$$A20_{ij} = \frac{A20_{i*j}A21_{ij}}{A21_{i*j}}.$$

■ Results

Table 2 lists the unweighted percentage of missing data for each recorded field on the Schedule A in tax year 1992. Though the number of missing forms is constant, the number of observed values for each field varies. For fields such as state tax refund, cash contributions, and total deductions, the amount of missing data is small. However, for fields which are not declared as often, such as medical expenses allowed and casualty loss, the potential amount of missing data can be significant. Even though one would expect most of the casualty loss imputed values to be zero, there exists a potential for a significant increase in the aggregate estimate. Missing data analysis is the only way to determine if the missing data cause a significant underestimation of the aggregate amounts.

The hot-deck imputation procedure described in the methodology section was implemented on the 1992 PIT Sample for California resident returns only. Table 3 lists the changes from the pre-imputation to the post-imputation weighted estimates of the number and dollar amounts for all recorded fields on the Form 1040 Schedule A in the form of the percent differences. For the majority of the fields, the percent increase from the pre- and post-estimates was at or below 1%. Since the lower AGI classes had a larger amount of missing data, the changes in the

TABLE 2: Maximum Possible Percentages of Missing Data for the Form 1040 Schedule A by AGI Class

Field	AGI Class					
	Below $100,000	$100,000 to $150,000	$150,000 to $300,000	$300,000 to $500,000	$500,000 to $2,000,000	Over to $2,000,000
Medical Expenses	2.04	3.43	2.27	0.86	1.09	0.58
Medical Expenses	3.34	21.73	22.91	16.98	28.57	25.00
State Tax Refund	1.08	1.08	0.79	0.28	0.32	0.16
Real Estate Tax	1.09	1.15	0.84	0.29	0.33	0.17
Other Taxes	1.15	1.20	0.93	0.34	0.41	0.21
Mortgage Interest Reported	1.17	1.24	0.94	0.34	0.42	0.24
Mortgage Interest Not Reported	10.34	8.55	6.50	2.33	3.35	2.25
Deductible Points	5.18	3.96	2.80	1.05	1.42	0.92
Investment Interest	7.63	8.11	3.37	0.87	0.73	0.30
Cash Contributions	1.10	1.14	0.83	0.29	0.33	0.17
Noncash Contributions	1.96	1.70	1.41	0.55	0.72	0.41
Contribution Carryover	11.97	56.67	36.42	14.59	12.14	4.67
Total Contributions	1.22	1.12	0.82	0.29	0.33	0.17
Casualty Loss	58.85	77.66	69.23	50.00	58.62	41.67
Moving Expenses	47.84	32.62	26.81	15.79	24.11	21.74
Business Expenses	3.98	3.29	2.88	1.42	2.28	1.73
Other Expenses	1.87	1.95	1.37	0.48	0.53	0.25
Total Expenses	1.49	1.66	1.24	0.45	0.50	0.24
Expenses Allowed	2.66	3.80	3.41	1.78	2.35	1.54
Misc. Deductions	24.23	27.77	20.00	7.22	7.02	3.65
Total Deductions	1.12	1.22	0.83	0.29	0.32	0.16

estimates were greater. As expected, the fields with a large potential for significant changes in the estimates had mostly zeros imputed, and their estimates only increased marginally. Similar results were found for the other supporting schedules.

■ Conclusions and Further Research

The hot-deck imputation procedures developed for the different supporting schedules are a good first step in the development of data augmentation procedures for the PIT Sample. Though increases in the aggregate estimates for the 1992 data were small, there is little reason to expect whether this will be true for any subsequent year. The development and refinement of the imputation procedures will continue rather than exploring other missing data techniques, since the PIT Sample has multiple purposes and estimates based on imputation are easily and readily available. In the future, the FTB Research Bureau will enhance the current imputation procedures, develop more advanced models (such as the multi stage regression imputation) for comparison or replacement, extend imputation procedures to the other supporting forms, impute itemized deductions observations for standard deduction filers, develop imputation procedures for the Bank and Corporation sample, and develop variance estimates for certain income items.

Enhancements to the current imputation procedures can be accomplished in two ways. First, changes to the adjustment cells may have an effect

TABLE 3: Percent Increase from Pre-Imputation to Post-Imputation Estimates for the Form 1040 Schedule A by AGI Class

Field		AGI Class					
		Below $100,000	$100,000 to $150,000	$150,000 to $300,000	$300,000 to $500,000	$500,000 to $2,000,000	Over to $2,000,000
Medical Expenses	Number	0.76	0.14	0.08	0.03	0.00	0.00
	Amount	0.86	0.29	0.20	0.12	0.00	0.00
Medical Expenses Allowed	Number	1.44	1.06	1.09	0.64	0.00	0.00
	Amount	0.93	0.16	0.12	0.06	0.00	0.00
State Tax Refund	Number	0.92	1.07	0.79	0.28	0.30	0.16
	Amount	0.83	1.07	0.73	0.34	0.30	0.13
Real Estate Tax	Number	0.89	1.14	0.81	0.30	0.32	0.17
	Amount	0.81	1.06	0.75	0.28	0.32	0.09
Other Taxes	Number	0.98	1.09	0.70	0.30	0.31	0.17
	Amount	0.81	1.21	0.70	0.32	0.28	0.27
Mortgage Interest Reported	Number	0.91	1.13	0.86	0.28	0.33	0.14
	Amount	0.73	1.03	0.81	0.21	0.35	0.14
Mortgage Interest not Reported	Number	1.10	1.08	0.45	0.34	0.29	0.00
	Amount	0.95	0.83	0.47	0.65	0.51	0.00
Deductible Points	Number	0.75	1.17	0.78	0.20	0.36	0.18
	Amount	0.47	0.87	1.24	0.01	0.25	0.06
Investment Interest	Number	1.40	0.95	0.35	0.37	0.24	0.12
	Amount	0.32	0.21	0.02	0.08	0.08	0.01
Cash Contributions	Number	0.93	1.10	0.80	0.28	0.30	0.17
	Amount	0.76	1.06	0.60	0.16	0.13	0.06
Noncash Contributions	Number	0.88	1.05	0.69	0.31	0.28	0.16
	Amount	0.52	0.67	0.44	0.14	0.15	0.01
Contribution Carryover	Number	0.30	0.00	1.23	0.00	0.00	0.00
	Amount	0.00	0.00	0.20	0.00	0.00	0.00
Total Contributions	Number	0.94	1.09	0.80	0.28	0.31	0.17
	Amount	0.76	1.01	0.63	0.21	0.15	0.05
Casualty Loss	Number	0.00	0.00	0.00	3.85	0.00	0.00
	Amount	0.00	0.00	0.00	0.24	0.00	0.00
Moving Expenses	Number	0.00	1.20	1.75	0.00	0.00	0.00
	Amount	0.00	0.79	1.87	0.00	0.00	0.00
Business Expenses	Number	0.79	0.59	0.14	0.22	0.17	0.35
	Amount	0.60	0.80	0.23	0.18	0.22	0.02
Other Expenses	Number	0.54	0.37	0.09	0.07	0.08	0.10
	Amount	1.03	0.44	0.07	0.25	0.25	0.35
Total Expenses	Number	0.44	0.35	0.12	0.07	0.09	0.10
	Amount	0.61	0.64	0.19	0.22	0.24	0.31
Expenses Allowed	Number	0.88	0.81	0.34	0.28	0.36	0.62
	Amount	0.61	0.67	0.19	0.24	0.19	0.46
Misc. Deductions	Number	0.68	1.30	0.00	0.32	0.32	0.00
	Amount	0.14	2.41	0.00	0.01	2.37	0.00
Total Deductions	Number	1.34	1.24	0.86	0.30	0.32	0.16
	Amount	0.98	1.12	0.74	0.26	0.29	0.11

on the estimates. The procedure for choosing the nearest neighbor is being reviewed. Currently, the data are sorted by one covariate and the previous return is taken. This method was used for convenience and ease of programming. It would be better to define a distance equation for one or more covariates and use this measurement to find the nearest neighbor. The programming is more difficult, and it is not exactly clear how to define the distance measure.

More advanced imputation procedures could be developed. Early on, a two-stage imputation procedure based on mean or regression imputation was considered. However, the complexity of programming and the preanalysis necessary to implement this imputation technique made the procedure difficult to implement in a production setting. The first stage attempted to model the probability of a zero value in each adjustment cell using a Bernoulli distribution. The probability of success was estimated using the ratio of (number of times the variable was nonzero) by (number of times the form on which the variable is present was nonempty). The second stage used either the mean of the adjustment cell or a regression predictive value. In both cases, the values were scaled so that the total from the supporting schedule matched the value on the main form. The first stage of the imputation procedure did not work well, since it treated each imputation as an independent trial, so that the covariance structure was not preserved.

Also, Tobit and Logit regression models were considered for the first stage. However, because of the complexities encountered, it was deemed necessary to conduct further studies before implementation.

The development of imputation procedures for the other PIT supporting forms, the Bank and Corporation sample, and the standard deduction filers needs to be developed. Imputation procedures similar to the existing imputation procedures should be created for consistency. The development of an imputation procedure to impute itemized deduction line items for all standard deduction filers would be useful for the PIT Model. But two problems arise. First, only the standard deduction amount of the filer is known, not the amount to which the itemized deduction would total. Second, there are no observed values on which to base the imputation. In 1987, one member of the Research Bureau developed a mean imputation procedure based on adjustment cells for some of the common itemized deductions (e.g., state income taxes, mortgage interest, and cash contributions). This methodology is being studied and may be implemented into the PIT Sample in the future. Variance estimates of the aggregate amounts have not been widely reported by the Research Bureau. However, variance estimates and coefficients of variation are computed for various income items based on the two-stage stratified random sample variance [3]. These estimates are used to examine the accuracy of the sample, so that changes can be made to subsequent years, if necessary. Since single imputation procedures generally lead to the underestimation of the variance, a modification must be made to the variance estimates [4]. A jackknife variance estimate for hot-deck imputation procedures may be applicable to the imputation procedures developed, but further research is necessary.

■ Notes and References

[1] Chamberlain and Spilberg (1991), *The California Personal Income Tax Micro-simulation Model*, Occasional Paper Series -- California Franchise Tax Board (FTB/OPS 91-03).

[2] Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley, lists several different imputation procedures and is a good reference for other missing data techniques.

[3] Cochran (1977), *Sampling Techniques*, New York: Wiley.

[4] A proposed technique by Rao and Shao appears in "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," *Biometrika*, 1993. ■