THE CPS HOT DECK: AN EVALUATION USING IRS RECORDS

Martin David and Robert Triest, U.S. Bureau of the Census and University of Wisconsin--Madison

1. INTRODUCTION

Imputation for missing values has been a thorny problem for agencies producing multi-purpose surveys subject to secondary analysis by persons other than the data collectors. Use of matching methods that, in effect, duplicate values on some donor records to supply the missing items has lead to considerable comment (Welniak and Coder, 1980) culminating in a sharp exchange between Lillard, Smith, and Welch (1982) and the Population Division of the Bureau of the Census concerning the appropriateness of the matching method, the "Hot Deck" (HD), used to supply values for missing wages and salaries in the March income supplement to the Current Population Survey (CPS).

This analysis probes weaknesses of the hot deck by exploiting validating information in the IRS matched to 1981 CPS income supplement data. The imputations analyzed in this paper are generated from the logarithmic and the ratio models discussed in Little and Samuhel (1983). Largely for convenience, the modeling effort began with a ten percent random sample of the CPS. It was apparent that the hot deck could draw from a much larger population of observations to supply missing data. Thus any comparison between model-based imputations and the hot deck were asymmetric. To overcome that asymmetry, the hot deck algorithms were applied to the information contained in a sample of one-tenth of all the households. The outcome of that imputation is termed 1/10 CPS.

The analyses below review the performance of six imputation methods--the present CPS hot deck (CPS), 1/10 CPS, a logarithmic model with and without empirical residuals, and a wage rate (Ratio) model with and without empirical residuals. Comparisons are made to IRS tax return reports of wages and salaries and adjustments of that value which allow for response error, conceptual differences in the CPS and IRS measures, and mismatching.

The comparisons are predicated on matching CPS to IRS data. The matched records are a selection of the population with predictable characteristics. Selected records include a greater proportion of middle-aged persons than young or old. Selection favors middle and upper-income persons and discriminates against the poor, and particularly against poor transients and illegal immigrants. The effect of this selection on the comparisons is not evaluated in this paper. The comparisons do represent a preponderance of all wage and salary workers and are significant for that reason.

A second caveat must be raised about the comparisons. Wage and salary imputations is a multivariate problem. Information on the recipiency of wages and salaries may be missing; ancillary information on hours and weeks worked and occupation may also be missing. The comparisons presented indicate the performance of models for the amount of wages and salaries, conditional on the availability of recipiency and other ancillary information. We report on imputations in which only amounts were missing. However, this is a peculiar subsample of the cases with missing data. Therefore, for the remaining cases we present a hybrid imputation in which hot deck values imputed for recipiency are assumed to be real and model estimates are computed on the imputed ancillary values. As a result we can say how modeling may affect amounts of imputations for all missing data cases, but we must remain agnostic as to whether modeling could improve the joint imputation of amounts and ancillary data.

2. NOTATION AND COMPARISON VALUES

y_i represents measured CPS wages and salaries for respondents $i \in R$. z_i denotes imputed values for non-respondents $i \in N$, m_i denotes the comparison, of "true value" of wages and salaries for i = 1, ... N.

Then

$$y_i - m_i = \text{observational error}$$

$$z_i - m_i = \text{imputation error}$$

We measure bias of the imputation method by both absolute $(\bar{z} - \bar{m})$ and relative $(\bar{z} / \bar{m})$ measures. The dispersion of error is gauged by mean absolute error $(\overline{|z - m|})$ and mean relative absolute error $(\overline{|z/m - 1|})$.

To evaluate imputations, the comparison value m_i must be free of observational error, otherwise understanding of imputation error will be confounded. If IRS wages and salaries (V_i) are used as a proxy for m_i, two polar positions can be asserted: (a) $V_i = m_i$ and any difference $y_i - V_i$ is observational error or (b) $y_i = m_i$ and any difference $y_i - V_i$ is associated with differences in the IRS and CPS measurement system. These two points or view are important because $y_i \neq V_i$ in general. Regressing CPS wages and salaries on IRS wages and salaries and the square of the IRS amount, leads to rejection of the hypothesis that there is a simple linear relationship between the CPS and IRS amounts.

$$Y_i = 1.026\ V_i - .2109 \times 10^{-5}\ V_i^2 \qquad (1)$$

$$(.0088) \qquad (.232 \times 10^{-5})$$

$$N=4437 \qquad R^2 = .93$$

The quadratic term is significant at the 99 percent level. The quadratic relationship between CPS and IRS wages and salary estimated in regression (1) indicates that CPS and IRS wages are approximately equal at low earnings levels, but CPS earnings average only 86 percent of IRS earnings at an IRS earnings level of $80,000.

This nonlinear relationship might indicate a problem of increasingly severe underreporting of CPS wages as the level of earnings increases; an alternative explanation is that undetected mismatches in the sample are causing regression towards the mean and force a downward (more negative) bias in the coefficient of the quadratic term. To investigate the latter hypothesis, regressions were estimated for observations which are from IRS joint returns. Since the Social Security numbers of both spouses must be matched to the same return for joint filers, mismatches are much less likely for this return type. If the quadratic relationship were due to mismatches, the magnitude of the quadratic coefficient would be smaller when the regressions are estimated on only the joint filers. The quadratic relationship is actually more pronounced for the joint filers. This supports the hypothesis of increasing underreporting of wages to the CPS as IRS earnings rise.

A part of the differences modelled above reflects coverage of IRS and CPS wage amounts. It is certainly possible for the CPS report to exceed or fall short of the IRS amount depending on the circumstances surrounding receipt of wages. For example, sick pay may be included in the IRS total and excluded from the CPS report. Tips may be fully reported to CPS and unreported to IRS. The nature of the difference between CPS and IRS undoubtedly depends on the level of income. More cannot be said without a careful reinterview to establish the sources of reported CPS income and reconcile them to tax return data. The function estimated represents a "gross" relationship between y and m, and is certainly similar to tabulations by Herriot and Spiers (1975) and Kilss and Alvey (1976) based on the 1972 Exact match between IRS and CPS.

The two points of view, (a) and (b) above, imply different comparison values. If IRS values are truth, they can be used as a proxy for m_i. Alternatively, if CPS methods yield true reports, IRS values must be adjusted by a relationship, such as (1), which transforms V_i to an unbiassed estimate Y_i^e or the report obtained by CPS methods. Both comparison values will be used.

3. COMPARISON OF IMPUTATION METHODS

Overall imputed values for missing data do not fall far short of the IRS comparison values. The first row of Table 1, Part A shows the ratio of the sum of CPS hot deck imputed wages and salaries to the sum of IRS wages and salaries for the same persons. Imputed values fall short of the IRS total by 3.0 percent for single filers and 11.2 percent for joint returns. To ascribe meaning to this shortfall is perilous, since comparison of imputations to IRS values adjusted for response error yields a quite different picture: no shortfall for the separate returns and a 3.7% shortfall for the joint returns.

The original data used to estimate our parametric models show approximately the same degree of completeness as the imputed CPS values in the case of single filers (last row, Table 1A). A similar comparison shown for the joint filers is less informative as only cases where both members of a couple have complete data are included.

The comparison of imputation methods yields few surprises. The 1/10 CPS hot deck appears to impute somewhat more income, but achieves this completeness at the expense of substantially increased mean absolute relative error as can be seen by comparing columns 2 and 4 for the CPS and 1/10 CPS methods. Model predictions over the whole sample yield results very close to the full hot deck for the ratio of aggregate imputations to IRS values. The mean relative absolute error is smaller than the corresponding value for the hot deck, reflecting the fact that a portion of the variance is suppressed in each cell when only conditional expectations are used to generate the imputation. When residuals are chosen from the sample of observed data and are added to the predicted values, the mean relative absolute error rises above the full CPS hot deck for both types of models.

However, it is comparison between the 1/10 CPS and the model which is probably most pertinent, since these methods are based on the same sample of respondent information. The full CPS includes substantially more observed data. The relative mean absolute error for the models is slightly smaller than for the 1/10 CPS. This finding suggests that the modelling has been able to capture all the important features of the hot deck imputation.

Adjustment of the IRS comparison value brings imputations for both single and joint return imputations within five percent of the comparison value. Unfortunately, the adjustment increases the mean absolute relative error by about ten percent, reflecting an increase in error for the persons with low IRS values.

In Table 1B this interpretation is confirmed by computation of mean absolute error rather than relative values. The absolute errors shown highlight the fact that comparisons for single and joint returns are significantly different. The mean absolute error for the joint returns tends to be more than two times the mean absolute error of separate returns. A variety of influences are operative. Two effects tend to make the mean absolute error of couples less than twice that of separate filers: Spouses may have zero wages. Also, the mean absolute error reflects the sum of two imputations with a random component; the sum will therefore reflect less variability if the random errors are independently chosen or are negatively correlated. One tendency in the opposite direction can be cited. Couples tend to have characteristics that place them higher in the distribution of earnings than single persons. One might expect a larger absolute error for that reason.

In dollar terms the mean absolute error of comparisons involving adjusted IRS values is about 10 percent smaller than the unadjusted IRS. This reflects the number of persons for whom $y_i < V_i$ as IRS rises above $40,000.

The mean difference between imputations and the IRS comparison value makes it clear that error of the imputations, in the aggregate, is extremely small for all imputation methods when the adjusted IRS income is used for a comparison.

Performance of the several imputation methods is detailed according to the amount of missing information in Table 2. In the first two entries of column 1 we see almost no difference between reported values and the IRS aggregate within the respondent sample. (Joint returns shown include only those for which no data were missing for either spouse.) The columns of the table are ordered in the direction of increasing amounts of missing data. Column 2 reflects situations where only one amount must be imputed, but recipiency is known. For joint filers, in column 2, the earnings or lack of earnings for one spouse must be known while only the amount is missing for the other. In column 3 both recipiency and amounts are unknown for one person. The remaining columns reflect a situation in which amounts are unknown for two persons; column 4 shows cases where information as to recipiency is complete. Column 5 shows cases where recipiency is not known for one person. Column 6 shows the case where recipiency is not known for either marital partner.

The most curious aspect of Table 2 is that the aggregate of imputations falls substantially below the IRS aggregate in the cases where something is known about the wages and recipiency of one marriage partner (columns 2 and 3). This deficiency does not appear to be a shortcoming in the models alone. The full hot deck is 14 to 19 percent below the IRS aggregate, which is approximately the same shortfall as for the models. Part of the difficulty is attributable to the model. For observed data, predicted values overestimate earnings when both persons receive wages and underestimate earnings when there is only one spouse earning. This finding reflects a deficiency in the models, namely the omission of the predictor variable "labor force status of the spouse" which is present in the hot deck matrix. Including this variable in the models should eliminate part of the difference between 1 and 2 earners for the respondents, and improve the model-based imputations for nonrespondents.

Adjusted IRS comparison values in Table 2 lend credibility to this conjecture. The somewhat low values of model imputations for cases where there is only one-earner (or one potential earner) on joint returns are balanced by some over imputation when both persons are imputed (cf. col. 3 and 6). Mean absolute error varies by twenty to twenty-five percent across types of imputations.

The clearest evidence of the relative advantages of modelling in comparison to the hot deck is displayed in Table 3. Imputations are classified according the level at which imputation is made in the full CPS hot deck. Level 1 contains the most elaborate matrix of adjustment cells; level 3 contains the most minimal matrix. The classification does not capture the fact that eight percent of all matches are made outside of the original group in the full CPS hot deck. Those cases could be matched at level 1, but information is lost as some of the observed data are suppressed.

Table 3 clearly reveals the level 1 imputations by the full hot deck as the most precise, and a pattern of increasing relative absolute error as adjustment cells are deleted in the process of moving to levels 2 and 3. The 1/10 hot deck has a larger mean relative absolute error and deteriorates to a greater extent than the full hot deck. Neither model-based method deteriorates to the same degree as the 1/10 hot deck in level 3; the increased absolute error may be related to the number of persons for whom information is lost in the hot deck algorithm when matching within the group is no longer possible. (Most persons matched at level 2 will be matched within the group for the 1/10 hot deck, which may explain why that entry lies between the errors shown for the full hot deck levels 2 and 3.) Predictions from the model appear to have relatively less error in comparison to hot deck methods at levels 2 and 3, which is what we would suspect because level 1 is the level at which detailed occupational categories are used. The degree to which mean relative absolute errors of the logarithmic model at level 2 and level 3 approach those of the full hot deck is some measure of the success of the model in capturing important effects.

A second aspect of the performance of the models is captured in Table 4. The table displays the ratio of the mean absolute error for imputed observations to the mean absolute error for the predicted values for the observed data from which the model was estimated. It seems important to note that the excess of that ratio above one is easily explained by mismatching of data for predicted values less than $10,000. Above $10,000, the prediction errors for nonrespondents and respondents are comparable for single filers, but are markedly greater for nonrespondents than for respondents among joint filers. Two explanations can be offered: (1) The distribution of non-respondents differs from respondents, after effects of covariates have been removed; (2) the relatively small sample of persons with wages above $40,000 implies that it is difficult to detect and parameterize differences in higher and lower wage workers. The solution to the second problem is to increase sample size on which parameters affecting high wage and salary cases are based.

Our comparison excludes cases where weeks and hours are imputed. Inclusion of those cases raises the ratios that appear in the entries of Table 4 so that the excess of error for missing cases above the error for cases with observed data is relatively uniform over values of the predicted wage level. We believe that the implication is that weeks and hours imputed by the hot deck do not closely reflect actual work experience. It is possible that explicit modelling of a censored distribution using Tobit, or similar methods, could offer a substantial gain because the method relates work effort to a latent dimension and imposes more structure on the data.

4. CONCLUSIONS

A. Non-respondents are Different from Respondents

Those who have worried about the naive imputation implied by assuming that non-response can be ignored have good cause to worry. Imputation of the grand mean from the respondent sample to the non-respondents would cause substantial bias.

B. Modelling is an effective imputation method.

In the many comparisons that we have made, modelling appears to have slightly less mean absolute error than the most comparable hot deck, the 1/10 CPS hot deck procedure. The relatively uncomplicated model that we used to generate values limits the number of interactions and provides significant smoothing for continuous variables such as weeks and hours worked. The model does not perform quite as well as the full hot deck, but the comparison is clearly unfair, as ten times more respondents are used to generate imputations in that procedure.

The modelling approach allows a ready transfer of empirical results from research and can be updated as easily as the hot deck.

C. Comparisons of CPS and IRS pose significant problems.

The exact match of the CPS to the IRS is both incomplete and subject to error. Incompleteness limits the value of these comparisons for some population groups, including some who have loss of income associated with family dissolution. Mismatching creates subtle biases as estimates of error for atypical populations will be contaminated by irrelevant comparison information. The approach of adjusting the comparison values taken in this paper reduces that problem, but other approaches, such as purposely introducing mismatches and extracting estimates from the changes in tabulated values also ought to be pursued.

Lastly, it is clear that aside from mismatching, there is a significant difference in response between IRS and CPS. Adjusting the comparison value to account for that response is essential before conclusions can be drawn about the inadequacies of imputation procedures.

D. None of the imputation methods show substantial bias

After IRS values have been adjusted, there is no evidence of systematic departures of imputed values from the comparison. This finding is perhaps the most significant, because it throws substantial doubt on the allegation that non-ignorable nonresponse is quantitatively important. It is true that the models do not impute values to non-respondents with as little error as they mimic observed values for respondents; that difference remains to be carefully explored and understood.

E. Improvements in modelling require multivariate methods.

The most clear deficiency of the modelling approach lies in its failure to condition on earnings or lack of earnings of the spouse. The hot deck incorporates some data in that direction, but we believe that more extensive modelling, such as Betson and Van Der Gaag (1982) would be even more revealing, as the hot deck does not fully reproduce differences in the mean earnings of one- and two-earner couples.

The greatest limitation of the modelling approach is the lack of a clear methodology for handling the multi-variate problem that is resolved by present hot deck methods. Modelling the joint distribution of wages and salaries recipiency, weeks and hours worked, and self-employment income requires strategic choices that are not easy. The scanty evidence presented here suggests that some gain in modelling weeks and hours may be possible, and that considering the joint distribution of spouses wages and salaries is important.

ACKNOWLEDGEMENTS

This paper reports research conducted by the authors with Roderick Little and Michael Samuhel under the American Statistical Association / Census Bureau Fellowship program, supported by grants from the National Science Foundation and the U.S. Bureau of the Census. The authors gratefully acknowledge the advice and guidance of John Coder, Gordon Green and Chuck Nelson of the Population Division of the Census Bureau, and useful comments on an earlier draft by Lawrence Ernst.

DISCLAIMER

All work involving the March 1981 CPS and the 1980 individual income tax records in the development and subsequent analysis of the the matched file was done by employees of the Bureau of the Census to preserve the confidentiality of the CPS respondents. No one other than Census Bureau employees has access to this file. The only products of this study are statistical tabulations summarizing the results of the analysis.

REFERENCES

Betson, D. and Van der Gaag, J. (1983), "Working married women and their impact on the distribution of welfare in the United States," Working paper, Institute of Research on Poverty, University of Wisconsin.

David, M., Little, R., Samuhel, M.,and Triest, R. (1983) "Nonrandom nonresponse models based on the propensity to respond" Proceedings of the Survey Research Section ASA.

Herriot, Roger, and Spiers, Emmett (1975), "Measuring Impact on Income Statistics between CPS and administrative sources", Proceedings Social Statistics Section ASA.

Kilss, Beth, and Alvey, Wendy (1976) "Further exploration of CPS - IRS - SSA Wage reporting differences in 1972" also in USHEW/ORS Report 11 op.cit. pp 57-78.

Little, R., and Samuhel, M. (1983) "Alternative models for CPS income imputation" Proceedings of the Survey Research Section ASA.

Welniak, E.J. and Coder, J.F. (1980), "A measure of the bias in the March CPS earnings imputation scheme, American Statistical Association 1983 Proceedings of the Section on Survey Research Methods, pp. 421-425.

Table 1
Comparison of Imputation Methods by Type of Return

Imputation Method and Comparison Method	A. Relative Comparisons*				B. Absolute Comparisons			
	Separate Returns		Joint Returns		Separate Returns		Joint Returns	
	Ratio: $\Sigma Z/\Sigma M$	Mean $\lvert Z/M - 1\rvert$	Ratio: $\Sigma Z/\Sigma M$	Mean $\lvert Z/M - 1\rvert$	Error	Mean absolute error	Error	Mean absolute error
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Hot Deck								
(a) Full CPS								
IRS	.970	.588	.888	.454	$100	$5700	$-2000	$11800
Adjusted IRS	1.000	.587	.963	.492	300	5400	-100	10500
(b) 1/10 CPS								
IRS	.967	.631	.926	.499	200	6200	-900	13100
Adjusted IRS	.997	.631	1.003	.541	500	6000	1000	11700
Logarithmic Model								
(a) Predictions								
IRS	.974	.479	.899	.400	100	4700	-1700	10500
Adjusted IRS	1.004	.474	.975	.434	400	4400	200	9100
(b) Plus residuals								
IRS	.988	.605	.875	.490	300	5900	-2300	12700
Adjusted IRS	1.018	.605	.949	.532	600	5600		11300
Ratio Model								
(a) Predictions								
IRS	.978	.478	.885	.396	200	4700	-2000	10500
Adjusted IRS	1.008	.474	.960	.429	500	4500		9000
(b) Plus residuals								
IRS	.968	.624	.887	.492	100	6000	-1900	12800
Adjusted IRS	.998	.625	.962	.533	400	5800	-0	11500
Sample Size	2915		3076		2915		3076	
Observed Data for Model								
IRS	.995	.114	.976	.155				
Adjusted IRS	1.009	.157	1.020	.119				
Sample Size	1823		1974					

*A problem with zero weights has biased these numbers. The corrected values will be published with the expanded version of this paper in the Census Bureau Technical Report Series. Cases in which receipiency was imputed and the the IRS shows m_i = 0 (0.2%) appear to be the omitted group.

Table 2
Ratio of Imputations to IRS Comparison Values
($\Sigma Z/\Sigma M$) by Imputation Method, Type of Return
and Extent of Imputation

Comparison Method, Imputation Method and Return Type	Extent of Imputation					
	No Imputations (1)	One Amount (2)	One Receipt (3)	Two Amounts (4)	One Amount Plus One Receipt and Amount (5)	Two Amounts Plus Two Recipiencies (6)
			UNADJUSTED IRS			
Hot Deck Full CPS						
Single	1.00**	1.02	.94	*	*	*
Joint	.98**	.86	.81	.98	.87	.92
Log-Model Plus Residuals						
Single	.99	1.00	.98	*	*	*
Joint	.96	.82	.79	.95	.98	.94
Ratio-Model Plus Residuals						
Single	1.00	.98	.96	*	*	*
Joint	.96	.82	.79	.97	1.02	.97
			ADJUSTED IRS			
Hot Deck Full CPS						
Single	1.01	1.03	.98	-	-	-
Joint	1.02	.96	.88	1.04	.94	.98
Log-Model Plus Residuals						
Single	1.01	1.02	1.02	-	-	-
Joint	1.00	.91	.86	1.01	1.06	1.00
Ratio-Model Plus Residuals						
Single	1.01	.99	1.00	-	-	-
Joint	1.00	.92	.86	1.03	1.10	1.03
Sample Size						
Single	1823	1105	1810	*	*	*
Joint	1974	862	900	449	118	747

* = Logically impossible
** = Observed CPS values in numerator

Table 3
Mean Relative Absolute Error
by Imputation Level

Imputation method	Level of Imputation Within Hot Deck Group						All
	Level 1		Level 2		Level 3		
	Error	Ratio*	Error	Ratio*	Error	Ratio*	
Hot Deck							
(a) full CPS	.467	100	.490	105	.539	115	.488
(b) 1/10 CPS	.516	110	.517	111	.619	133	.533
Logarithmic Model							
(a) predicted	.421	90	.406	87	.458	98	.421
(b) plus residuals	.517	111	.502	107	.574	123	.520
Ratio Model							
(a) predicted	.417	89	.401	86	.461	99	.417
(b) plus residuals	.512	110	.525	112	.564	121	.526
Sample size	2525		2545		983		

*Error divided by Error for full hot deck level 1, times 100.

Table 4
Ratio of mean absolute error for missing
data to observed data $\overline{|Z - M|}^{\text{missing}} / \overline{|Y - M|}^{\text{observed}}$
by model used to impute wages and type of return
($\overline{|Z - M|}$ restricted to cases where
only wages and salary amounts are missing)

Predicted Value of Wages ($000's)	Logarithmic Model		Ratio Model	
	Single	Joint	Single	Joint
<10	0.89	0.93	0.87	1.12
10-20	0.97	1.31**	0.97	1.20**
20-30*	-	1.32	-	1.50
30-40	-	1.71	-	1.54
40-50	-	1.66***	-	1.91***
50-75	-	1.36	-	1.45
All	1.09	1.77	1.03	1.81

*Calculations are not reported for cells where either the numerator or the denominator has less than 20 observations.
**Numerator calculated from just 21 observations.
***Both the numerator and the denominator where calculated from less than 50 observations.