W1

8404

UNIVERSITY OF WISCONSIN-MADISON

SSRI Workshop Series

PREDICTING WAGES AND SALARIES: A NEW PERSPECTIVE FROM THE CURRENT POPULATION SURVEY

M. David
R. Little
M. Samuhel
R. Triest

8404

SOCIAL SYSTEMS RESEARCH INSTITUTE

Social Systems Research Institute

University of Wisconsin

PREDICTING WAGES AND SALARIES: A NEW PERSPECTIVE FROM THE CURRENT POPULATION SURVEY

M. David
R. Little
M. Samuhel
R. Triest

8404

January 1984

PREDICTING WAGES AND SALARIES: A NEW PERSPECTIVE FROM THE POPULATION SURVEY

1. Motivation

In the last fifteen years data on the sources of income reported in the March income supplement to the Current Population Survey have become increasingly complex. What began as a simple set of questions asked in sequence has become an interlocking network of questions that more precisely identify income sources, particularly from government programs. At the same time resistance to answering questions has increased. The result has been an increase in the proportion of income and the proportion of the population for whom some item of income is missing. See Table 1.

The potential for non-random non-response is large, and must be explicitly or implicitly addressed in any estimation using the data. The Bureau of the Census has sought to minimize the problems of analyzing observed data by simulating values for the missing cases. The simulated values are incorporated in micro-data sets and are flagged as imputed values. To be acceptable, an imputation must satisfy several obvious statistical properties:

A. The imputed value should be unbiased.

B. The distribution of imputed values should approximate the distribution of the true values which are missing. That is, imputations should not distort the true distribution by heaping imputed values at particular points or by falling outside of logical bounds for the missing data.

C. Estimates of covariances, relationships between variables should be preserved in the imputed data.

D. Valid estimates of standard errors should be obtained from the imputed data.

Imputations are based on observed data. Each imputed data point constitutes a forecast of values in the unobserved population. Understanding the adequacy of imputations according to the four criteria above, constitutes an understanding of a particularly rigorous forecasting process in which attention must be given to both the conditional expectations and the stochastic properties of the forecast. Forecasting in this sense is the means by which statistically-matched synthetic data sets are created. Inadequacies of imputation procedures are likely to have analogues in such matched data sets.

2. Wage and salary (WS) imputation in CPS

The Bureau of the Census uses an algorithm to impute missing values in CPS that is known as the hot deck (HD). The procedure partitions all observed data into a matrix of cells based on sixteen dimensions. The same partition is applied to cases in which data are missing. Within each cell values for the missing datum (or data) are transferred from an observed case, who becomes a donor, to the missing case, who becomes the recipient. As the order in which cases are assigned to the donor and recipient matrices contains a chance element, the imputation has stochastic character, although matching is not a process of independent random selections.

A number of questions have been raised concerning the adequacy of the HD as a procedure for imputing missing data. The most strident criticism (Lillard, Smith and Welch, 1982) is that the probability of reporting income or wages and salaries depends on income so that response is nonignorable (Rubin, 1976). The implication is that the HD leads to biassed estimates of the missing values and that computations based on imputed data will not give desired parameter estimates.

We chose to test this proposition directly. We model WS using simple regression techniques. Forecasts from the models are then tested against values reported to the IRS on tax returns. The resulting comparisons give insight into two issues: (1) whether explicit modelling techniques can improve upon the HD as a forecasting device and (2) whether either or both techniques give rise to substantial bias that makes the maintained assumption of ignorable non-response untenable.

The procedure used in the HD can be made clearer by the following illustration. Let the indices i,j, and k reference categories of three classifying variables. The value of any observed wage and salary in the cell (i,j,k) is given by the identity:

$$y_{ijk\ell} \equiv \mu_{ijk} + \varepsilon_{ijk\ell} \qquad (1)$$

where ℓ is an index running over individuals, μ_{ijk} is the mean for the cell, and $\varepsilon_{ijk\ell}$ is the residual associated with the ℓth observation. When a non-respondent m is matched to a respondent ℓ in this cell, the imputed hot deck value may be thought of as

$$y^{HD}_{ijk\ell} = \mu_{ijk} + \varepsilon_{ijkm} \qquad (2)$$

The HD fails to provide an imputed value when the class (i,j,k) is empty in the responding sample, when (2) is the imputation rule. Since the HD matrix has about 8 x 10^9 cells and 113,000 observations, only about one-third of the 17,000 cases of missing data are imputed using the full detail.

To assure that an imputed value <u>always</u> exists, the classification is collapsed by deleting dimensions or aggregating categories in a particular dimension whenever no donor exists with the identical characteristics as the non-respondent.

Several features of the HD algorithm should be noted. No restriction is placed on the imputation process that would tend to "smooth" the response surface implied by the μ_{ijk} in adjacent cells. Similarly, no restrictions are placed on the stochastic processes generating residuals within each cell. The use of imputed values in conjunction with observed values reduces the bias of estimated means, but it decreases the estimated variance and covariance. These features led us to explore a modelling procedure that replaces (1) with a structured model.

3. A forecasting model for imputations

The earnings identity is central to our approach to modelling for imputation:

$$y_{i\ell} \equiv w_{i\ell} \cdot \text{WKS}_{i\ell} \cdot \text{HRS}_{i\ell} \qquad \begin{array}{l} i = H,S \\ \ell = 1\ldots L \end{array} \qquad (3)$$

Where y denotes wages and salary, w is the wage rate and WKS and HRS are weeks and hours worked per week by the i^{th} marriage partner in the ℓ^{th} household. Thus i has only two values, householder and spouse. In general we anticipate that earnings of the householder and spouse are not distributed independently, and that the level of work effort for any individual is conditioned by the wage rate. This line of thinking gives rise to structural models of earnings such as that in Betson and van der Gaag (1983).

Because we are interested in comparing a modelling approach to the HD, and because the HD imputes $y_{i\ell}$ after weeks, hours, and occupation have been imputed, we chose to model earnings conditionally on known values for WKS, HRS, and occupation. Two forms for an estimating equation were considered:

$$y_{i\ell} = F(WKS_{i\ell}, HRS_{i\ell}, x_{i\ell}) \cdot \varepsilon_{i\ell} \qquad (4.1)$$

$$w_{i\ell} = G(WKS_{i\ell}, HRS_{i\ell}, x_{i\ell}) + \eta_{i\ell} \qquad (4.2)$$

where $x_{i\ell}$ denotes a set of predictors for individuals who received WS. Preliminary testing revealed that the relationships for householders and spouses were statistically indistinguishable. Data were pooled across all earners within a household.

The form of this problem differs from a typical labor supply analysis because of the nature of the non-response problem. People are more willing to supply information on occupation, weeks and hours worked than they are to supply information on the level of earnings (and its source, i.e. wages and salaries, self-employment earnings, and farm income). Indeed, some persons report the receipt of earnings by type, but do not report the amounts. Because non-reporting may be related to income level, it appears useful to explore the capacity of this conditional relationship to forecast WS when positive evidence of eanings exists in the conditioning variables measuring weeks, hours, and occupation. We are able to study the error of this forecast because records for the persons with missing wage and salary information are matched to IRS records of earnings.

4. Structure of the Data

To understand the contribution of this analysis and its relationship to earlier work, we must explain the nature of that match and the correspondence between CPS and IRS measures of wages and salaries. The entire CPS for March 1981 was matched to IRS income tax returns for 1980 using reported social security numbers.[1] The following notation is useful:

$y_{i\ell}$ wages and salaries, CPS definition $i = H,S$; $\ell = 1 \ . \ . \ .L$

m_ℓ wages and salaries, IRS definition for (joint) tax return filing units

$c_{i\ell} = 1$ if CPS $y_{i\ell}$ is reported, $c_{i\ell} = 0$ otherwise

$s_{i\ell} = 1$ if IRS m_ℓ is matched to CPS, $s_{i\ell} = 0$ otherwise.

This notation makes clear that the population falls into four cells with the corresponding information

$$\begin{array}{rll} & & \text{Cell} \\ (s_{i\ell}, c_{i\ell}, y_{i\ell}, m_\ell) & = (1,1,y_{i\ell}, m_\ell) & \text{I} \\ & = (1,0, \ / \ , m_\ell) & \text{II} \\ & = (0,1,y_{ij}, \ / \ ) & \text{III} \\ & = (0,0, \ / \ , \ / \ ) & \text{IV} \end{array}$$

where / designates a missing value.[2]

Our procedure will be to estimate the parameters for (4) over cells I and III. Forecast values of WS, $y^Q_{i\ell}$, can be generated for all cells. Traditional measures of goodness of fit pertain to I and III. The comparisons that generate insight into the selection problem come from cell II, where $y_{i\ell}$ can be compared to m_ℓ.

The notation reflects one troublesome feature of the IRS comparison values. The amount reported is the amount of earnings received jointly by the husband and wife when joint returns are filed. Therefore CPS forecasts must be aggregated over householder and spouse whenever they are matched to a joint return. For convenience we define $Y_{\ell} = y_{H\ell} + y_{s\ell}$ when returns are filed jointly and $Y_{\ell} = y_{i\ell}$ otherwise.

A match to the IRS data may fail for a number of reasons, the most obvious being failure of the respondent to report a Social Security number. However, incorrect Social Security numbers may be reported, and that can lead to mismatches, particularly with single returns. Among joint returns, we restrict attention to those returns on which both marriage partners report filing the same type of tax return and the amount of m is identical on both records. About 17 percent of records were rejected as likely mismatches before continuing the analysis.

After imposing the foregoing consistency requirement, the probability of matching IRS to CPS data was studied using logistic regression. The results conform to our expectations. Old people, young people, blacks, and households with a low income have lower than average probability of being matched. The estimated probability of being matched varies widely. A middle-aged white person with some post-secondary education who lives in a household with more than $15,000 of income, has a 79 percent probability of being matched. In contrast a black person less than 20 years old who lives in a household with less than $10,000 annual income has a probability of being matched of only 14 percent percent. See Table 2. It seems apparent that the use of the IRS to validate forecasts of earnings will be more representative for some population groups than for others. Fortunately, matching seems relatively complete for persons likely to receive large WS.

Unfortunately, m_{ℓ} is not identical to Y_{ℓ}, because of conceptual differences and reporting errors. This leads us to calculate a calibration function

$Y_\ell = H(m_\ell, x_{H\ell}) + \eta_\ell$. Only cell I can be used in this calculation. It would be possible in principle to correct the parameter estimates for the selection involved in the match using selection models such as Heckman (1976). However, such methods are heavily dependent on model assumptions and were not tried here. We report comparisons over cell II for both $Y_\ell^Q - m_\ell$ and for $Y_\ell^Q - H(\cdot)$. The latter form of the comparison differences the forecast income and the estimated reported income, which is generally smaller than the difference between forecast income and the IRS comparison value.

The function H(•) was estimated by OLS as

$$\underset{(.0088)}{Y = 1.026m} - \underset{(.232 \times 10^{-6})}{.2109 \times 10^{-5} m^2} \qquad n = 4437 \qquad \bar{R}^2 = .93 \qquad (5)$$

(Standard errors in parentheses). The equation indicates that Y and m are approximately equal at low earnings levels, but CPS WS reports average only .86 of IRS at an IRS earnings level of $80,000.[3] This relationship is pooled across return types for which no significant difference in the relationship could be determined. The fact that the quadratic effect estimated for joint returns proved more negative than for individual filers justifies our belief that the relationship is a response problem, not a mismatching problem, as we are certain that the rate of mismatching is lower in the joint filing population than among the single filers (due to the consistency of information required between spouses).

5. Relationship to earlier work

Earlier work on the problem of imputed income was of considerably narrower scope than the present investigation. Greenlees, Reece and Zieschang (1982), hereafter GRZ, estimated a model for m based on cell I and studied the characteristics of m^Q forecast for cell II. Thus GRZ do not make use of all cases where y is observed and estimate the parameters of their model conditional on a match to the IRS, i.e., s = 1. In addition, GRZ limit the universe of interest to married men (with non-working spouses) who worked for the full year in non-agricultural employments and who filed joint tax returns. (This population was approximately one-fifth of the population that we study.) These population re-strictions obviate difficulties associated with the difference in separate and joint filing of tax returns, but clearly result in in a limited and idiosyn-cratic universe for study. By estimating a forecast relationship on m, GRZ avoids the problems of comparing IRS and CPS WS amounts, which were quantitified in (5).

Lillard, Smith, and Welch (1982), hereafter LSW, estimate a model for y from cells I and III, and report on the implications for y^Q in cells II and IV that can be derived from the maintained distributional assumptions that underly their model of the selection problem. They conclude that the HD grossly understates values of WS. LSW limit analysis to civilian white males between the ages 16 and 65 and exclude persons who have self-employment earnings or whose major activity is schooling.

In contrast our forecast is generated from a model that is estimated on all persons who receive more than $100 in WS. Men and women are included without regard to marital status, other sources of earned income, or the extent of weeks and hours worked. All three of these studies model earnings conditionally on reports of the receipt of wage and salary amounts, about which we will need to say more later.

6. Characteristics of the model

Two functions are served by discussing the specification of the model.

First we establish variables omitted from the specifications used by GRZ and LSW. Second we indicate the extent to which our model restricts the implicit specification of the HD and provides smoothing for the resultant response surface. LSW include only region (south), schooling, and experience in their specification for the earnings relationship. Inspection of Table 3 reveals the much more extensive specification of our model. (Of course, their selection of a smaller population implies some interactions that we did not specify.) GRZ use a similar specification to LSW, but add a simple classification based on occupation and industry at the 1-digit level.

The function F(•) is estimated by transforming (4.1) into logarithms and then estimating parameters, a typical approach and that used by GRZ and one of several used by LSW. A quadratic form in ln(WKS) and ln(HRS) is included in the specification of F. The form (4.2) estimates an equation for a wage rate, which is again often specified in labor supply work; earnings are derived from (4.2) using the identity in (3). Thus the difference in the specifications of F and G relate to two mathematical properties -- (a) the logarithmic transformation of F gives less weight to observations with large earnings than the additive form G used in (4.2); (b) the specification F tests for non-linearities in the relationship between WKS, HRS, and y. The motivation for the non-linear specification is to permit differences in the wage rate implied for part-time and part-year workers as against full-time or multiple job-holding earners.

The major difference in the specification of both equations from common econometric practice was the inclusion of a set of 64 categories to estimate effects of particular occupation-industry combinations. The detail in this dimension was motivated by a desire to mimic detail included in the HD, at least to a degree. The most detailed occupation-industry classification used in the HD includes about 375 categories. As LSW observe, some of this detail is important in distinguishing earners with high wages. Even the HD is unable to exploit such detail as only about one-third of imputations are made without collapsing occupational detail. The collapsed occupational codes include 47 categories that precisely correspond to detailed groupings in published tables. LSW's analysis suggests that some aggregation in those categories ranges across groups with widely different wages (see LSW Table 11).[4] We therefore adopted a larger set of categories that distinguishes high- and low-skill groups within the professional and managerial groups. The categories were chosen, in part, by identifying outliers in a regression fit with the 47-category grouping based on a sample of the data different than the sample used below.

The continuous treatment of several variables implies a smoothing of effects across categories used in the HD matrix. Age, education, weeks and hours worked, and age of the spouse are all treated as continuous variates. Except for spouse's age, all these effects are significantly non-linear, as indicated by the significant quadratic terms. See Table 3.

The most important findings in the log regression are the highly significant quadratic terms in the WKS and HRS, and the statistical importance of the occupational classification in a relationship that includes considerable flexibility in functional form for variables for human capital and age. The importance of the occupational detail applies at the level of the 47 groups used by the Census, and is not simply an artifact of the pretesting that we did to establish the additional categories. The usual models of earnings that rest on separate equations for wage rates (independent of work effort) and work effort imply that (3) can be used to generate estimates of earnings.

The quadratic terms in the logarithmic model imply that the identity in (3) does not imply a wage rate that is independent of WKS and HRS.

The third finding, which comes as no surprise to most workers on earnings, is that the residuals from the regression model for the logarithm of wages and salary are heteroskedastic and asymmetric. In particular, the mean square residual (MSR) declines as the level of earnings rises:[5]

$$\underset{(.0694)}{\text{MSR} = 0.2786} - \underset{(.00776)}{.068}(y - 8.87) \qquad \overline{R}^2 = .011 \quad n = 6970 \tag{6}$$

This apparently small relationship is important because the conditional expectation for WS from the log model for individual i is given by:

$$\hat{y} = \exp\ (\hat{b}'x + (\hat{\sigma}^2/2)(1 - x_i'(X'X)^{-1}x_i)) \tag{7}$$

where $\hat{b}$ are the estimated coefficients of the log model and $\hat{\sigma}$ is the estimated standard error of the residuals and X is the design matrix and x_i is the vector of predictors for individual i. We therefore replace $\hat{\sigma}$ with MSR from (6) to obtain the unbiassed conditional expectation for the earnings of workers with characteristics x_i.

The skewness of the distribution of residuals might be exploited to obtain adjustments for nonrandom nonresponse, as is in the maximum likelihood estimates of GRZ and LSW based on stochastic censoring models. Such methods place strong reliance on the assumption that the residuals are symmetric in the unselected population, an assumption which cannot be tested.

Modelling the wage rate was somewhat more straightforward as the functional specification gives an unbiassed conditional expectation by multiplying the conditional forecast from the wage relationship by WKS times HRS. The modelling of the wage rate also includes significant effects for weeks and hours worked, corroborating the quadratic effects in the log regression. Occupation makes a larger relative contribution to the explained sum of squares. This is understandable since much of the variance in earnings is associated with differences in WKS and HRS and variance is attenuated by computing w from (3).

Modelling wage rates also gives somewhat more weight to high wage workers, because the logarithmic transformation of earnings tends to weight those with minimal labor market activity heavily. However, forecast values from the wage rate equation can be negative, and we dealt with that problem by replacing negative forecasts with zeroes in the few cases where this problem arose.

No conceptual or estimation gains appear to be obtained from using (4.1) in preference to (4.2) so we continued to use forecasts from both in making comparisons to IRS values.

7. Modelling the response propensity

In addition to the predictors shown in Table 3, we also fit an extended model including functions of the estimated propensities to respond as predictors. These models are fitted by first calculating a probit on $c_{i\ell}$ across cells I - IV. The regressors are variables which survey methodology indicates will be related to the willingness and ability of the respondent to recall the earnings amount (Cannell and Henson, 1974) and which earlier researchers found to be significant. Our principal innovation was to include non-response to a categorical question on income asked early in the question sequence as an indicator of defensive attitude towards reporting income items in general. Prior work is confirmed in Table 4 which clearly indicates that missing categorical data on the control card is a harbinger of later propensity not to respond to the earnings question.

Several models can be invoked to make use of the estimated $\hat{c}_{i\ell}$ the probability of response, obtained from the probit. (David, et al. 1983). One possibility is that respondents who share characteristics of those with a high propensity not to respond will also have behavior more similar to non-respondents than persons who do not share those characteristics. This line of thinking leads to a specification of forecasting models in which interactions are permitted between the parameters of the model and the probability of response. In the extreme, there may be no basis for pooling observations across groups with different probabilities to respond. A second line of thinking, familiar to labor economists, is that a correlation between residuals in the probit equation and earnings equation can be exploited through the Mills ratio to give unbiassed estimates for the parameters of the earnings model, if the correlation results from the selection process and the random variables have a joint normal distrubtion. This is the IV technique first suggested by Hechman (1976) and discussed by Olsen (1980).

We studied the effect of adding functions of $\hat{c}_{i\ell}$ to the model. A ten-category classification based on levels of $\hat{c}_{i\ell}$, the linear predictor from the probit, $b'x_{i\ell}$, and the Mills ratio were added, in separate regression experiments. In no case did the specification yield a significant contribution to the regression, and parameter estimates were unaffected. This is in marked contrast to the results of LSW and GRZ. Further work proceeded without including $\hat{c}_{i\ell}$ in the regression.

The principal conclusion to be drawn from these investigations is negative. We have been unable to locate the significant effects of the propensity to respond that were detected by GSZ and LSW in their work with more limited populations and more limited models of earnings. Much of the effect discovered by them appears to be the result of specification errors in the model for earnings. The corollary is that LSW's assertions about the bias of HD is likely to be unfounded.

We offer a different kind of evidence on that assertion by comparing forecast values from the model estimated on cells I and III to the IRS-based comparison values in cell II.

8. Adding residuals to the conditional forecast

A unique problem imposed by the requirement that simulated values have the same distribution as the population, is that some means must be found to supply random variation in the forecasts that is orthogonal to the conditional expectation. The most obvious procedure is to take random drawings from a normal distribution $N(0,\hat{\sigma}^2)$. This will not work both because of the heteroskadasticity which we already reported and because of deviations from normality.

To find a residual to add to the conditional expectation, we exploited the idea underlying the HD. Namely, actual residuals in the observed population were assigned to records where earnings was missing. The selection of residuals was controlled by classifying both respondents and non-respondents according to $\hat{y}$, and then selecting the residual from respondents within the same class as the non-respondent. The classes were defined by intervals of $2000 and residuals were assigned using a single random start and a systematic selection across the range of $\hat{y}$ included in the interval. This procedure showed less deviation from the true distribution in cells I and III than alternatives, and did not fail to be reliable at any level of $\hat{y}$.[6]

9. Comparisons of forecast values with m for the non-respondents

To this point in our report earnings data have come from respondents. The value of the match to IRS is that the tax record supplies m for non-respondents. Denote the forecast value by z. Differences in the forecast values and the IRS value are reported as average error $(\overline{z} - \overline{m})$, average relative error $(\overline{z}/\overline{m})$, average absolute error $|\overline{z - m}|$, and average relative absolute error $|\overline{z/m-1}|$. Actual values, m, from IRS are replaced by H(m), equation 5, in tables showing comparison to adjusted IRS values.

The entire CPS file was used to provide non-respondents for comparisons. A 10% sample of respondents was used to fit (4). Therefore model-based comparisons do not have the same donor information as the HD. One of the comparisons generates HD forecasts based on an identical respondent sample - we term this the 1/10 HD.

Our model-based methods concern only the imputation of WS amounts, given given ancillary information on occupation, recipiency of wages and salary and hours and weeks worked. In practice some or all of this ancillary information is sometimes missing. Rather than restrict comparisons to the peculiar subsample of cases with only WS amount missing, we included all cases with WS amounts missing and used HD estimates of missing ancillary variables. As a result we can assess how the modelling of amounts works on all non-respondents, but we cannot tell whether a multivariate model could improve the joint imputation of amounts and ancillary data. We reluctantly must leave this as a topic for future research.

Overall, imputed values for missing data do not fall far short of the IRS comparison values. The first row of Table 5 shows the ratio for 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 the shortfall is perilous; comparison of imputations to IRS values adjusted for response error indicate no shortfall for the seperate 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 5A). A similar comparison for the joint filers is less informative as only cases where both members of a couple have complete data are included.

A comparison of imputation methods in Table 5A yields only modest differences. The 1/10 CPS hot deck imputes somewhat more income than the full hot deck, but has 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 forecast. When residuals are chosen from the sample of observed data and are added to the predictd values, the mean relative absolute error rises above the full CPS hot deck for both types of models.

Table 6 displays the ratio of the mean absolute error for cell II to the mean absolute error for cell I, using the forecast in both cases. 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 nonrespondents differs from respondents, after effects of variates 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.

Theoretical considerations might suggest weaknesses of the forecast for small subgroups of the population, such as blacks. However, classifications of the performance of the imputation methods by socioeconomic subcategories of the population did not yield markedly different results from those we have presented for the whole sample. One further experiment can be reported. Following the lead of GRZ, we chose to model m, using a specification almost identical to Table 3. The chief differences are that GRZ's 9-category occupation scale was used, and the universe was limited to single filers. The regression was fit to cell I, and applied to cell II. The forecast error reported for the cases where WS amounts were missing, but recipiency had been reported, proved to be indistinguishable from the forecast error for observed cases.

This comparison obviates the problem of conceptual differences in the variables being compared which motivated the adjustment H(m) in our earlier tables. The lack of differences across cells I and II raises some doubt about the need for an elaborate correction for selection, such as that calculated by GRZ, but we did not have the time to replicate his results with the larger regression model.

10. Conclusions

For that part of the population for whom comparisons are possible, cell II, bias associated with the simulation of WS by models based on OLS appears to be on the same order of magnitude as the Census Bureau HD imputation procedure. The bias is not large relative to the uncertainty created by (i) conceptual differences in CPS and IRS values that are being compared and (ii) the likelihood of response errors associated with each source of data. Going beyond OLS estimation to an IV method for correcting for selection bias did nothing to improve the model, despite clear relationships between the added variables and the process of selection. This finding casts doubt on earlier work by LSW who found the IV method for correcting for selection bias to be significant in a subpopulation of the current universe, when using a much smaller number of regressors in their forecasting equation.

It would appear that several interpretations of these results are possible:

A. Non-response is ignorable, and forecasts from OLS models are satisfactory.

B. Non-response is non-ignorable, and biases in the selection must be understood by maintaining an assumption on the joint distribution of the selection process and earnings.

This investigation has been unable to lend support to B. The comparison of forecast values to cell II is not complete, as it is clear that a substantial selection is made in the process of matching CPS and IRS data. Being optimists we prefer interpretation A and conclude that reasonably simple weighting procedures or HD imputation result in satisfactory estimates from the CPS.

Table 1

Non-response and Allocation in the CPS March Income Supplement

		Number: Percent of Total		Amount
Year		Individuals	Families	Percent of Income
1979		—	--	18.9
1978		25.5	29.9	19.8
1976		17.1	23.9	17.4
1971	M	12.5		—
	F	9.3	14.6	
1965	M	10.9		—
	F	6.9	14.0	
1960	M	9.0		—
	F	6.0	10.5	

Source: Ono, M. (1971), p. 347
CPS P-60 (1976, 1978, 1979)

Table 2

Logit Analysis of the Probability of a Match*

Variable	Coefficient	Standard Error	Probability**
Constant	.81	.05	-
Age:			
< 20	-1.45	.07	-.36
65 +	-.75	.06	-.19
Black	-.54	.07	-.14
Control Card Income:			
Not coded	-.77	.07	-.19
$0-10000	-.95	.06	-.24
$15000 +	.20	.05	.05
Post Secondary Education	.28	.04	.07
Marital Status:			
Single	-.23	.04	-.06
Married with Absent Spouse	-3.07	.19	-.77

Sample Size = 29,248

*The dependent variable equals one if the CPS observation is matched to an IRS record, zero otherwise. All independent variables are also binary. Estimation was by likelihood maximization.

**The partial derivative of the probability of a match with respect to the independent variable evaluated at the 50% probability level.

$$P = \frac{e^{xB}}{(1+e^{xB})}; \qquad \frac{\partial P}{\partial X_i} = B_i \cdot \frac{e^{xB}}{(1+e^{xB})^2};$$

$\frac{\partial P}{\partial X_i} = .25B_i$ when $P = .5$ and

$\frac{\partial P}{\partial X_i} = .19B_i$ when $P = .75$)

Table 3 (Part 1)

Statistics for the Parameters of the Wages and Salaries Model

Variable	Description	Coefficient	Significance Level of t-Statistics
LNHRS	Natural log of hours worked per week	-.01878	.0661
LNWKS	Natural log of weeks worked per year	.02799	.0023
LNWSQ	Square of LNWKS	.00303	.0093
LNHSQ	Square of LNHRS	.00867	.0000
LNWLNH	Interaction between LNHRS and LNWKS	.01632	.0000
NEAS	North East Region	-.0046	.0175
NCEN	North Central Region	-.00313	.0762
STH	Southern Region	-.00869	.0000
SMSA1	SMSA size 3,000,000	.02038	.0000
SMSA2	SMSA size 1,000,000-3,000,000	.01566	.0000
SMSA3	SMSA size 250,000-1,000,000	.00916	.0000
SELFEMPL	Recipiency of self-employment income	-.05015	.0000
AGE	Age in years	.00459	.0000
GRADE	Highest grade attended	.00193	.2149
AGED	Interaction of AGE and GRADE	.3147x10^{-4}	.0488
AGSQ	Square of AGE	-.5290	.0000
EDSQ	Square of GRADE	.1576x10^{-4}	.7499
SAGE	Age of Spouse	.1787x10^{-3}	.3090
SAGSQ	Square of SAGE	-.05092	.1174
EVM	Ever married	.01929	.0000
NEVM	Never married	-.00563	.1337
ROTH	Race other than black, white	-.164x10^{-3}	.9662
RBLK	Black race	-.0136	.0001
SEXF	Female	-.01361	.0000
EVMSEXF	Interaction of SEXF and EVM	-.02405	.0000
BLKF	Interaction of RBLK and SEXF	.01360	.0034
SRF2	Spouse self-respondent	.00465	.0035
CCMISS	Household income missing in the Basic CPS	-.01724	.004
(Constant)		.3605	.0000
OCC1-OCC63	Occupation/Industry Recodes (see Table 3) -part 2-	(See Table 3) -part 2-	(See Table 3) -part 2-

Table 3. Occupation/Industry Dummy Variables Included in Wages and Salary Model
(Part 2)

Variable	Profession, technical, and kindred workers	1980 Census Occupation Code	1980 Census Industry Code	Coefficient	Significance Level of t-statistics
OCC 1	Accountants	001	-	.01399	.0303
OCC 2	Computer specialists	003-005	-	.02614	.0072
OCC 3	Engineers	006-023	-	.04413	.0000
OCC 47	Architects, Lawyers, Judges	002, 030-031	-	.03588	.0143
OCC 4	Scientists & mathematical specialists	034-054	-	.03394	.0012
OCC 5	Chiropractor, Dentists, Optometrists, Podiatrists, Health Practioners n.e.c.	061-073 exc. 64, 65, 72	-	-.21701	.0000
OCC 48	Pharmacists, Physicians, Veternarians	64, 65, 72	-	.01612	.0906
OCC 9	Health Workers	74-85 exc. 75	-	01180	.1003
OCC 49	Registered Nurses	75	-	.02809	.0000
OCC 50	Religious Workers	86, 90	-	-.07504	.0000
OCC 6	Teachers, college & university	102-140	-	.02050	.01444
OCC 7	Teachers, elementary and high school	141-145 exc. 143	-	.00332	.4503
OCC 55	Prekindergarten and kindergarten	143	-	-.01708	.2588
OCC 8	Engineering and Science Technicians	150-162	-	.01217	.0434
OCC 51	Designers, Editors, Reporters, Musicians and Composers	183-185	-	.01879	.0177
OCC 10	Other professions	All Other Between 001-195	-	.00836	.0562
	Managers and administrators except farm				
OCC 11	Managers, manufacturing	201-245	107-398	.03522	.0000
OCC 12	Managers, retail trade	201-245 exc. 216, 230	607-698	-.00122	.8065
OCC 52	Managers and superintendents, building	216	607-698	-.0616	.0000
OCC 56	Managers, restaurants cafeterias, bars	230	607-698	-.02291	.0109
OCC 13	Managers, finance, real estate insurance and wholesale trade	201-245	507-588, 707-718	.026	.0000
OCC 14	Managers, public administration	201-245	907-937	.02025	.0066
OCC 15	Other managers, n.e.c.	201-245	All other industry codes	.01670	.0001
	Sales				
OCC 54	Hucksters, peddlers Newspaper, carriers, vendors	264, 266	-	-.08728	.0000
OCC 16	Sales, insurances stock agents and brokers	265, 271	-	.00626	.5059
OCC 53	Sales, real estate	270	-	-.05126	.0000

Variable	Profession, technical, and kindred workers	1980 Census Occupation Code	1980 Census Industry Code	Coefficient	Significance Level of t-statistics
OCC 17	Sales, wholesale trade & manufacturing	281,282	-	.01948	.0017
OCC 18	Sales, retail sales clerks	283	-	-.02565	.0000
OCC 19	Salesmen, retail trade	284	-	-.00532	.5714
OCC 20	Sales, other n.e.c.	all other between 260-285	-	-.02491	.0205
	CLERICAL				
OCC 21	Clerical, bookkeepers Clerical, cashiers & counter	305	-	.7857-3	.874
OCC 22	Clerks exc. food	310,314	-	-.01638	.0004
OCC 57	Enumerators, interviewers, teachers aids	320,382	-	-.04764	.0000
OCC 23	Clerical, office machine operators	341-355	-	-.00705	.2834
OCC 24	Clerical, secretaries and stenographers	370-372, 376	-	.00129	.7360
OCC 25	Clerical, typist	391	-	-.00666	.2694
OCC 26	Clerical other n.e.c.	all other between 301-395		reference	category
	Crafts and kindred workers				
OCC 27	Crafts, blue collar supervisors,	441	-	.01556	.3227
OCC 28	Crafts, construction	all between 401-575 exc. 441,430,473	067-077	.01067	.0172
OCC 58	Electricians	430	067-077	-.01427	.0385

Variable	Profession, technical, and kindred workers	1980 Census Occupation Code	1980 Census Industry Code	Coefficient	Significance Level of t-statistics
OCC 29	Crafts, mechanics & repairmen n.e.c.	470-495, exc. 473	all except 067-077	.00790	.1098
OCC 59	Automobile mechanics	473	all except 067-077	.04480	.0000
OCC 30	Crafts, metal crafts, exc. mechanics	403, 404, 442, 446,454, 461, 502-504,514, 533,535, 536, 540,561,562	all except 067-077	.01816	.0018
OCC 31	Crafts, other n.e.c.	all other between 401-575	all except 067-077	.01121	.0024
	Operatives				
OCC 32	Operatives, durable goods manufacturing	601-695, exc. 640,663	107-259	.0081	.0249
OCC 33	Operatives, nondurable goods manufacturing	601-695, exc. 640,663	268-398	.137-3	.9747
OCC 60	Mine operatives, n.e.c.	640	-	.04297	.0000
OCC 63	Sewers and stitchers	663	-	-.01677	.0227
OCC 35	Operatives, transport	701-715	-	-.0061	.1031
	Nonfarm Laborers				
OCC 36	Laborers, construction	740-785	107-398	-.00348	.5873
OCC 37	Laborers, manufacturing	740-785	107-398	.3353-3	.9638
OCC 38	Laborers, other industries	740-785	all except 067-077, 107-398	-.0137	.0008

Variable	Profession, technical, and kindred workers	1980 Census Occupation Code	1980 Census Industry Code	Coefficient	Significance Level of t-statistics
	Service Workers				
OCC 39	Service, private household	980-983	-	-.1111	.0000
OCC 63	Private household cleaners and servants	984	-	-.01677	.0227
OCC 40	Service, cleaning service	901-903	-	-.02704	.0000
OCC 41	Service, food service	910-916	-	-.03373	.0000
OCC 42	Service, health	921-926	-	-.01598	.0008
OCC 43	Service, personal	931-954	-	-.0323	.0000
OCC 44	Service, protective	960-964	-	-.02496	.0003
OCC 61	Police and detectives	964	-	.00626	.5059
OCC 45	Farmers and farm managers	801-802	-	-.09997	.0000
OCC 46	Farm laborers and supervisors	821-824	-	-.05508	.0000

$\bar{R}^2 = .811$ N = 6997

Table 4

Probit Analysis of Non-response (= 1)

Variable	Coefficient	T-ratio	Comment
Constant	-1.356	19.4	
Income not reported in categories on control card (=1)	1.773	19.1	
Self respondent (=1)	-.278	5.6	c
First year of income supplement (=1)	-.0924	2.1	b
Personal interview (=0)	.198	4.1	a,b
Census divisions (NE = 1, NC = 2, S = 3, W = 4)	-.0656	-3.2	a

Log likelihood = -1835

N = 7845

[a] Included in GRZ, corroborating effect.

[b] Included in LSW, corroborating effect.

[c] LSW include effects for reporting of child earnings or earnings or secondary family members. These are likely to be proxy reports and the effect is therefore in the same direction.

TABLE 5
Comparison of Imputation Methods by type of return

	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	Absolute error	Error	Absolute error
Imputation method and comparison method	(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					

Table 6

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	Logarithmic model		Ratio	
of wages ($000's)	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 were calculated from less than 50 observations.

FOOTNOTES

[1] All work involving the March 1981 CPS and the 1980 individual income tax records in the development and subsequent analysis of 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.

[2] One possibility which will not be discussed is that non-coverage in the CPS produces a situation where $c_{i\ell}$, $s_{i\ell}$ are unknown but the value of m_ℓ exists and can be determined.

[3] Similar relationships can be found in tabulations from the 1972 exact match CPS-IRS-SSA file in Herriott and Speirs (1975) and Kilss and Alvey (1980).

[4] It is curious that LSW ignore this problem in their model of WS which fails to include even 47 categories of occupational detail. In testing our specification, it is clear that the implied additive scaling for occupation that is reported in Table 3 is stable across sub-samples of the 1980 CPS. It aLso appears to perform relatively well in comparison to the HD in those cases where 3-digit occupational codes were used in the HD forecast.

[5] These empirical realities lead LSW to apply the Box-Cox transformation to the earnings data. Unfortunately, that resulted in an over-parameterization of their selection model.

(The mean logarithm of earnings is 8.87 and the standard error of the regression is .528.)

[6] This empirical selection of residuals has one great advantage over the HD as a method for simulating missing earnings values. When only one donor (or fewer donors than recipients) occurs, it is often necessary to replicate donor values for several recipients. In the modelling approach a single conditional expectation will be calculated, but residuals can be obtained from several donor observations. Variance of estimates from the model-based predictions will consequently be smaller than the variance of analogous HD estimates.

References

Afifi, A. A. and Elashoff, R. M. (1966), "Missing observations in multivariate statistics I: Review of the literature," Journal of the American Statistical Association, 61, 595-604.

Anderson, T. W. (1957), "Maximum likelihood estimates for a multivariate normal distribution when some observations are missing," Journal of the American Statistical Association, 52, 200-203.

Aziz, F., Kilss, B. and Scheuren, F. (1978), 1973 Current Population Survey-Administrative Record Exact Match File Codebook, Part I - Code Counts and Item Definitions, Washington, D.C., U. S. Department of Health, Education and Welfare.

Bailar, B. A., Bailey, L. and Corby, C. A., (1978), "A comparison of some adjustment and weighting procedures for survey data," in Survey Sampling and Measurement (Namboodiri, N.K., ed.), Academic Press: New York.

Beale, E.M.L. and Little, R.J.A. (1975), "Missing values in multivariate analysis," Journal of the Royal Statistical Society, Series B, 37, 129-146.

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 for Research on Poverty, University of Wisconsin.

Cannell, C. and Henson, R. (1974), "Incentives, Motives, and Response Bias," Annals of Economic and Social Measurement, 3, 307-318.

Chen, T. and Feinberg, S. E. (1974), "Two dimensional contingency tables with both completely and partially classified data," Biometrics, 30, 629-642.

Colledge, M. J., Johnson, J. H., Pare, R. and Sande, I. G. (1978), "Large scale imputation of survey data," Proceedings of the Survey Research Methods Section, American Statistical Association, 431-436.

Cook, R. D. (1977), "Detection of influential observations in regression," Technometrics, 19, 15-18.

David, M. R. and Triest, R. K. (1983), "The CPS hot deck: an evaluation using IRS records," Proceedings of the Survey Research Methods Section, American Statistical Association.

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, American Statistical Association.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B, 39, 1-38.

Ernst, L. R. (1978), "Weighting to adjust for partial nonresponse," Proceedings of the Survey Research Methods Section, American Statistical Association, 468-473.

Ford, B. N. (1976), "Missing data procedures a comparative study," Proceedings of the Social Statistics Section, American Statistical Association, 324-329.

Greenlees, W. S., Reece, J. S. and Zieschang, K. D. (1982), "Imputation of missing values when the probability of response depends on the variable being imputed," Journal of the American Statistical Association, 77, 251-261.

Haitovsky, Y. (1968), "Missing data in regression analysis," Journal of the Royal Statistical Society, Series B, 30, 67-81.

Hartley, H. O. (1958), "Maximum likelihood estimation from incomplete data," Biometrics, 14, 174-194.

Hartley, H. O. and Hocking, R. R. (1971), "The analysis of incomplete data," Biometrics, 27, 783-808.

Heckman, J. D. (1976), "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimation for such models," Annals of Economic and Social Measurement, 5, 475-492.

Herriot, R., and Speiers, E. (1975), "Measuring Impact on Income Statistics between CPS and administrative sources," Proceedings Social Statistics Section, American Statistical Association.

Herzog, T. N. and Rubin, D. B. (1983), "Using multiple imputations to handle nonresponse in sample surveys," to appear in Incomplete Data in Sample Surveys, Vol. 2, W. G. Madow, I. Olkin and D. B. Rubin, eds., New York: Academic Press.

Kalton, G. and Kasprzyk, D. (1982), "Imputing for missing survey responses," Proceedings of the Survey Research Methods Section, American Statistical Association.

Kalton, G. and Kish, L. (1981), "Two efficient random imputation procedures," Proceedings of the Survey Research Methods Section, American Statistical Association, 146-151.

Kilss, B. and Alvey, W. (1976), "Further exploration of CPS - IRS - SSA Wage reporting differences in 1972", USHEW/ORS Report 11, 57-78.

Lansing, J. B., Ginsberg, G. and Braten, K. (1961), "An Investigation of Response Error," Urbana: University of Illinois Press.

Lillard, L., Smith, J. P. and Welch, F. (1982), "What do we really know about wages: The importance of non-reporting and Census imputation," The Rand Corporation.

Lillard, L. A. and Willis, R. J. (1978), "Dynamic aspects of earnings mobility," Econometrica, 46, 985-1011.

Little, R.J.A. (1982), "Models for nonresponse in sample surveys," Journal of the American Statistical Association, 77, 237-250.

Little, R.J.A. and Samuhel, M.E.: "Alternative models for CPS Income Imputation," to appear in Proceedings of the Survey Research Methods Section, American Statistical Association 1983.

Olsen, R. J. (1980), "A Least Squares Correction for Selectivity Bias," Econometrica, 48, 1815-1820.

Oh, H. L. and Scheuren, F. (1980), "Estimating the variance impact of missing CPS income data," Proceedings of the Survey Research Methods, Section, American Statistical Association, 408-415.

Oh, H. L., Scheuren, F. and Nisselson, H. (1980), "Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data," Proceedings of the Survey Research Methods Section, American Statistical Association, 416-420.

Ono, Mitsuo, 1971. "Current Developments on Collecting Income Data in the CPS," Proceedings of the Section on Social Statistics. American Statistical Association, Washington, D.C.

Orchard, T. and Woodbury, M. A. (1972), "A missing information principle: theory and applications," Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 697-715.

Rubin, D. B. (1974), "Characterizing the estimation of parameters in incomplete data problems," Journal of the American Statistical Association, 69, 467-474.

Rubin, D. B. (1976), "Inference and missing data," Biometrika, 63, 581-592.

Scheuren, F. S. (1983), "Weighting adjustments for unit nonresponse." To appear in Incomplete Data in Sample Surveys, Vol. 2, W. G. Madow, I. Olkin and D. B. Rubin, eds., New York: Academic Press.

Schieber, S. J. (1978), "A comparison of three alternative techniques for allocating unreported social security income on the Survey of the Low-Income Aged and Disabled," Proceedings of the Survey Research Methods Section, American Statistical Association, 212-218.

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