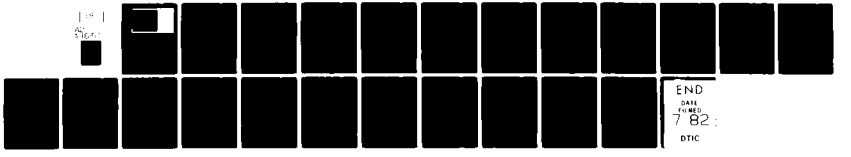
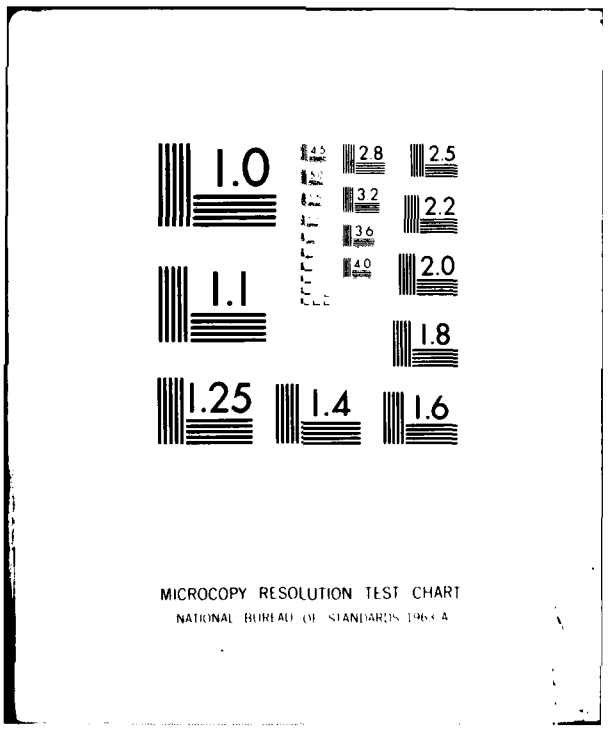
AD-A116 157 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 5/3
IMPUTING INCOME IN THE CPS: COMMENTS ON 'WHAT DO WE KNOW ABOUT --ETC(U)
APR 82 D B RUBIN DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2358 NL



END
DATE FILMED
7-82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

MRC Technical Summary Report #2358

IMPUTING INCOME IN THE CPS: COMMENTS ON "WHAT DO WE KNOW ABOUT WAGES: THE IMPORTANCE OF NON-REPORTING AND CENSUS IMPUTATION" BY LILLARD, SMITH AND WELCH

Donald B. Rubin

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

April 1982

Received January 5, 1981

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

IMPUTING INCOME IN THE CPS: COMMENTS ON "WHAT DO WE KNOW ABOUT WAGES: THE IMPORTANCE OF NON-REPORTING AND CENSUS IMPUTATION" BY LILLARD, SMITH AND WELCH

Donald B. Rubin

Technical Summary Report #2358
April 1982

ABSTRACT

Nonreporting of income in the Current Population Survey is an important problem affecting the many researchers using the data base. This paper discusses an approach to handling this problem proposed by Lillard, Smith and Welch, which applies selection models to a Box-Cox transformation of the income variable. Topics considered here include: the inadequacy of single imputation and the desirability of multiple imputation, the importance of the distinction between ignorable and nonignorable nonresponse, the sensitivity of inference to assumptions unassailable by the data at hand, and the possibility of using the CPS-SSA-IRS Exact Match File to study such assumptions. The Lillard, Smith and Welch paper accompanied by this discussion is to appear in a book presenting the proceedings of the NBER Labor Cost Conference to be published by the University of Chicago Press.

AMS(MOS) Subject Classifications: 6206, 6207, 62D05, 62P15, 62H99, 62P20

Key Words: sample surveys, nonresponse, missing data, Box-Cox transformations, selection models.

Work unit Number 4 - Probability and Statistics

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

IMPUTING INCOME IN THE CPS: COMMENTS ON "WHAT DO WE KNOW ABOUT WAGES: THE IMPORTANCE OF NON-REPORTING AND CENSUS IMPUTATION" BY LILLARD, SMITH AND WELCH

Donald B. Rubin

1. Introduction.

"What do we know about wages: the importance of non-reporting and census imputation" by Lillard, Smith and Welch (LSW) is a very interesting study of nonresponse on income items in the Census Bureau's Current Population Survey (CPS). LSW points out that the CPS is a major source of income data for economic research even thought the nonresponse rate on income items is about 15% - 20%. This level of nonreporting of income, especially if concentrated among special types of individuals, should be of substantial concern to researchers in economics. As emphasized in LSW, however, most published economic research ignores this problem when using CPS data. The major reason that researchers can ignore this problem is that before CPS public-use tapes are released, the Census Bureau imputes (i.e., fills in) missing income data (as well as other data). Although imputed data are flagged to distinguish them from real data, it is evidently easy for researchers to be seduced into ignoring this distinction and treating all values, imputed and real, on the same basis.

LSW is divided into three main sections. In the first, facts are presented concerning the CPS, income nonrespondents, and the procedure used by the Census Bureau to impute (i.e., the "hot deck"). In the second section, a statistical model is formulated to explain income

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

nonresponse, specifically, a selection model using Box-Cox transformations to normality. The third section summarizes empirical results obtained when the selection model is applied to CPS data.

My discussion of LSW will roughly follow the outline in LSW with digressions and extensions. My sections do not, however, follow in one-to-one correspondence with theirs. After characterizing income nonreporters in Section 2 and describing the Census Bureau's hot deck procedure in Section 3, in Section 4 I point out the need for multiple imputation if uncertainty due to nonresponse is to be properly reflected in an imputed data set. Section 5 provides definitions of ignorable and nonignorable nonresponse, while Section 6 describes the LSW selection model and emphasizes that external information is needed to justify the acceptance of the LSW model or any other particular model for nonresponse as an accurate reflection of reality. Finally, Section 7 briefly describes the CPS-SSA-IRS Exact Match File, which might be used to help provide such external information.

2. Who are the Nonrespondents on Income Questions?

Of central importance for determining whether the 15% - 20% nonresponse rate on income questions is of major concern is the extent to which income nonreporters are different from income reporters. If the nonreporters were just a simple random sample from the population of reporters and nonreporters, the loss in efficiency of estimation created by ignoring the nonreporters altogether would be of little concern.

There is a great deal of evidence, however, showing that nonreporters do differ from reporters in important ways. One such piece of evidence that LSW presents is especially interesting. Apparently, if we were to plot "probability of nonresponse on income items" vs. "amount of actual income", the relationship would be U-shaped: moderate nonresponse at low incomes, low nonresponse at moderate incomes and very high nonresponse at high incomes. Moreover, LSW's evidence suggests that this U-shaped relationship is created by the existence of two primary types of income nonreporters. The first type is called "general nonreporters" because they have a high nonresponse rate on many CPS questions, not just income questions. These people tend to have low incomes and approach CPS questions in a generally reluctant manner. The second type of income nonreporter is called "specific nonreporters" because on most CPS questions, that is non-income questions, they have low nonresponse rates, whereas on income questions their nonresponse rates are very high (e.g., over 30%). The specific nonreporters tend to be professionals with high incomes, for example, doctors, lawyers, and dentists.

If we accept this interesting picture as relatively accurate, it seems to me natural and desirable to try to build a nonresponse model that explicitly recognizes the U-shaped relationship and the two types of income nonreporters. LSW, however, does not exploit this structure in its models, and instead uses a model for nonresponse asserting that conditional on some predictor variables (such as years of education), the relationship between probability of nonresponse on income items and income is monotonic. Of course one can criticize virtually any analysis for not fully exploiting some interesting features found in subsequent analyses. Consequently, my comment on this point should be viewed more as offering a suggestion for further study than as criticizing the work presented in LSW.

3. The Census Bureau's Hot Deck Imputation Scheme.

LSW provides an exceptionally clear and lucid discussion of the Census Bureau's procedure for imputation, the hot deck, which has been used since the early 1960's. The hot deck is a matching algorithm in the sense that for each nonrespondent, a respondent is found who matches the nonrespondent on variables that are measured for both. The variables used for the matching are all categorical, with varying numbers of levels (e.g., "gender" has two levels, "region of country" has four levels). If a match is not found, categories are collapased and variables are deleted so that coarser matches are allowed. Eventually, every nonrespondent finds a match; the matching respondent is often called (by hot deck aficionados) "the donor" because the donor's record of values is donated to the nonrespondent to fill in all missing values in the nonrespondent's record.

LSW points out that the number of variables used for matching and their level of detail has expanded over the years, and that imputed income can be sensitive to such rule changes. For example, between 1975 and 1976, years of education was added to the list of matching variables, and as a consequence, the imputed incomes of nonrespondents with many years of education increased substantially from 1975 to 1976. Such changes can create problems when comparing income data in different periods of time. A related problem is that even though the ideal match that is possible under the hot deck is closer now than it was years ago, many nonrespondents fail to find donors at this ideal level of detail. For one example, only 20% find donors in the same region of the country. For a second example, judges with ideal matches are imputed to earn approximately $30,000 more than judges without ideal matches.

The hot deck, by trying for exact multivariate categorical matches, is trying to control all higher order interactions among the matching variables. This task is very difficult with many matching variables when using a categorical matching rule, even if there is a large pool of potential matches for the nonrespondents. Related work on matching methods in observational studies investigates categorical matching methods and offers alternative matching methods (e.g., Cochran and Rubin, 1973; Rubin, 1976a, 1976b, 1980a; Rosenbaum and Rubin, 1981). I suspect that some of the more recent work (e.g., Rosenbaum and Rubin, 1981) may have useful suggestions for an improved hot-deck-like procedure. LSW does not suggest modifying the matching algorithm but rather suggests using explict statistical models.

4. The LSW Alternative to the Hot Deck and the Need for Multiple Imputation.

LSW suggests an alternative to hot deck imputation: (a) build an explicit model (described here in Section 6), (b) estimate the parameters of this model, and (c) impute by randomly drawing observations from this model with unknown parameters replaced by estimates. Before proceeding to describe the particulr model LSW uses, I have several general comments to make in this section and the next.

First, for the data producer, some form of imputation is almost required and often desirable even if not required. I believe the Bureau feels it cannot produce public-use files with blanks. Also, I believe it feels, and rightly so, that it knows more about the missing data than the typical user of public-use files. Furthermore, the typical user of public-use files will not have the statistical sophistication needed to routinely apply model-based methods for handling nonresponse, such as those reviewed by Little (1982). Of course, in any public-use file, all imputed values must be flagged to distinguish them from real values.

Second, imputation based on explicit modelling efforts may require much more work than implicit models such as the hot deck (or some other matching method for imputation) that can impute all missing variables at once no matter what the pattern of missing variables. Of course, this does not mean that explicit models should be avoided: explicit model-based methods are, in principle, the proper ones to handle nonresponse.

Third, when drawing values to impute, in order to obtain inferences with the correct variability, parameters of models must not be fixed at estimated values but must be drawn in such a way as to reflect uncertainty in their estimation.

Fourth, one imputation for each missing value, even if drawn according to the absolutely correct model, will lead to inferences that underestimate variability (e.g. underestimate standard errors).

Fifth, there exists a need to display sensitivity of answers to plausible models for the process that creates nonresponse since the observed data alone cannot determine which of a variety of models is correct.

These points are all leading to the suggestion to use multiple imputation as proposed in Rubin (1978a) and expanded upon in Rubin (1980b). Whether using an implicit model, such as the hot deck, or an explicit model such as employed in LSW, if imputation is used to handle nonresponse, multiple imputation is generally needed to reach the correct inference.

Multiple Imputation replaces each missing value by a pointer to a vector, say of length m, of possible values; the m values reflect uncertainty for the correct value. Imputing only one value can only be correct when there is no uncertainity, but if there were no uncertainty, the missing value would not be missing; consequently, multiple imputation rather than single imputation is needed when there are missing data.

The m possible values for each of the missing data result in m complete data sets, and these can be analyzed by standard complete-data methods to arrive at valid inferences. Suppose for example that the

m imputations were all made under one model for nonresponse, such as the LSW selection model, and suppose that with complete data we would form the estimate $\hat{Q}$ with associated standard error S. Let $\hat{Q}_i$ and S_i, $i = 1,\cdots,m$ be their values in each of the data sets created by multiple imputation. Then the resultant multiple imputation estimate is simply $\bar{Q} = \Sigma\hat{Q}_i/m$ with standard error $\sqrt{\Sigma(\hat{Q}_i - \bar{Q})^2/(m-1) + \Sigma S_i^2/m}$.

If the m imputations are from k different models, then those imputations under each model should be combined to form one inference under each model, and then the comparison across the k resulting inferences displays sensitivity of inference to the k different models.

5. The Distinction between Ignorable Nonresponse and Nonignorable Nonresponse.

Before introducing the LSW model and presenting its implications, I think that it is important to expand on the general issue of the kinds of models that can be built for survey nonresponse. Such models can be classified into ones with "ignorable" nonresponse and those with "nonignorable" nonresponse, the terminology being due to Rubin (1976c, 1978b). I believe that LSW's use of "random nonresponse" is intended to convey essentially the same notion, although I find the LSW use of this phrase somewhat inconsistent.

Under ignorable nonresponse models, respondents and nonrespondents that are exactly matched with respect to observed variables have the same distribution of missing variables. The Census Bureau hot deck operates under this assumption although it does not have to do so. For example, having found a donor for a nonrespondent, instead of imputing the donor's income, the hot deck algorithm could be instructed to impute the donor's income plus ten percent. If we accept the Census Bureau's hot deck as currently implemented, then we implicitly accept the hypothesis that nonresponse is ignorable, and then there is no need to be concerned with selection models, such as that used in LSW. Instead, under ignorable nonresponse, all energy should be focused on modelling the conditional distribution of missing variables given observed variables for respondents, since, by assumption, this conditional distribution is the same for nonrespondents and respondents. If missing values are to be replaced by imputed values, however, whether these values arise from implicit or explict models, a single imputation generally will underestimate variability. Consequently, the LSW statement

accepting the hot deck if operating at its most detailed level is not entirely appropriate if valid inferences are desired, even if nonresponse is ignorable.

Under nonignorable nonresponse models, respondents and nonrespondents perfectly matched on observed variables have different distributions on unobserved variables. The example of the modified hot deck which imputes donor's income plus ten percent is an implicit nonignorable nonresponse model; the LSW selection model is an explicit nonignorable model. When nonignorable nonresponse is possible, as with income nonreporting in the CPS, it is crucial to expose sensitivity of answers to different models, all of which are consistent with the data. An important contribution of the present paper is that it defines and illustrates the use of an expanded collection of such models.

Within the context of *imputation for missing* values, sensitivity to models can only be exposed through the use of multiple imputation, where for each missing value there are imputations under each model being considered (e.g., two imputations under the ignorable hot deck, two imputations under the nonignorable-(plus ten percent)-model, and two imputations under the LSW nonignorable selection model). Again, such multiple imputations are necessary in order to reach valid inferences under each model and to expose sensitivity of answers to population features not addressable by the observed data.

6. The LSW Nonignorable Model and Analysis.

Let Y be earnings, which is sometimes missing in the CPS, and let X be a vector of predictor variables (e.g., education, work experience), which evidently, is assumed to be always observed in the CPS. Define Y^* to be the Box-Cox (1964) transformed earnings ($Y^* = (Y^\theta - 1)/\theta$), Z to be an unobserved, hypothetical variable such that Y is missing if $Z > 0$, and suppose (Y^*, Z) given X is bivariate normal with correlation ρ.

If $\rho = 0$, nonresponse is ignorable, whereas if $\rho \neq 0$ nonresponse is nonignorable; as $|\rho| \to 1$, the extent of nonignorable nonresponse becomes more serious in the sense that the observed distribution of Y^* for respondents becomes less normal and more skewed. This defines the LSW model, and LSW obtains maximum likelihood estimates for all parameters, explicitly recognizing the truncation of Y at \$50,000 in the CPS. A quite similar model with $\theta = 0$ ($Y^* = \log(Y)$) is applied to CPS income data in Greenlees, Reece, and Zieschang (1982). The extension to other θ is certainly interesting and potentially quite useful. Of particular importance, it gives users a broader range of models for nonresponse to which sensitivity of estimation can be investigated.

It must not be forgotten, however, that the estimation of parameters is relying critically on the assumed normality of the regression of (Y^*, Z) on X: both θ and ρ are chosen by maximum likelihood to make the residuals in this regression look as normal as possible. If in the real world there is no (θ, ρ) that makes this regression like a normal linear regression, then there is no real reason to believe that the answers that are obtained by maximizing over θ and ρ lead to

better real world answers. A small artificial example I've used before (Rubin, 1978) illustrates this point in a simpler context:

> Suppose that we have a population of 1000 units, try to record a variable Z, but half of the units are nonrespondents. For the 500 respondents, the data look half-normal. Our objective is to know the mean of Z for all 1000 units. Now, if we believe that the nonrespondents are just like the respondents except for a completely random mechanism that deleted values (i.e., if we believe that mechanisms are ignorable), the mean of the respondents, that is, the mean of the half-normal distribution, is a plausible estimate of the mean for the 1000 units of the population. However, if we believe that the distribution of Z for the 1000 units in the population should look more or less normal, then a more reasonable estimate of the mean for the 1000 units would be the minimum observed value because units with Z values less than the mean refused to respond. Clearly, the data we have observed cannot distinguish between these two models except when coupled with prior assumptions. (p. 22)

Notwithstanding the above caveats, suppose we put our faith in the normal linear model for the bivariate regression of (Y^*,Z) on X. LSW produce some interesting empirical results using white males, 16-65 years old, in the 1970, 1975, 1976 and 1980 CPS. One interesting, but not surprising, result is that fixing θ at 1 ($Y^* = Y$) produces very different answers from fixing θ at 0($Y^* = \log(Y)$); if $\theta = 1$, nonrespondents are imputed to earn less than matching respondents, whereas if $\theta = 0$, nonrespondents are imputed to earn more than matching respondents. With θ fixed, the asymmetry in the Y^* given X residuals addresses the correlation ρ and so determines the extent to which the nonresponse is nonignorable. Thus, we have learned that the Y given X residuals are skewed left and the log(Y) given X residuals are skewed right. Further study shows that $\theta = .45$ provides a better fit to the data than either $\theta = 0$ or $\theta = 1$, but that the residuals are still skewed right: under $\theta = .45$ we find that nonrespondents are imputed to earn more than similar respondents: $\theta = .45$

leads to a 10% increase in average earnings over the CPS hot deck values, $18,000 vs. $16,000.

But we must remember that if the distribution of $Y^{(.45)}$ given X really has the right asymmetry that is observed when $Y^{(.45)}$ is regressed on X, then the adjustment created by assuming a selection effect on Z is entirely inappropriate, and (just as with the artificial half normal example) the data cannot distinguish between the ignorable and nonignorable alternatives. More precisely, suppose first that in the population, $Y^{(.45)}$ has a linear regression on X with a skew distirbution of residuals like that observed when we regress $Y^{(.45)}$ on X for the CPS data and that nonresponse is ignorable; such a model would generate data just like those we have observed, and then we should not be imputing higher incomes for nonrespondents than respondents with the same X values.

In contrast, suppose that $Y^{(.45)}$ in the population really has a normal linear regression on X and that the stochastic censoring implied by the LSW probit-nonresponse model in correct, i.e., nonresponse is nonignorable with this particular form; then as LSW shows, we should be imputing higher incomes for nonrespondents than respondents with the same X values. There is no way that the observed data can distinguish between these two alternatives; if the authors really believe Y^* given X in the population is normal for some θ, then they can correctly assert that the CPS hot deck procedure is biased. If they admit the possibility that Y^* given X is not normal or even symmetric for any θ, then they cannot legitimately assert that their answers are better than the CPS answers.

In the same vein, LSW's checking the accuracy of the LSW model by checking the prediction of respondents' values really does not adequately check the imputations of the model for nonrespondents. In particular, both the ignorable and nonignorable nonresponse models discussed above will accurately reproduce the observed data for respondents, but will give very different results for nonrespondents. In order to address which model is more appropriate, we really need data from nonrespondents or some external information about the distribution of reported incomes in the entire population.

7. The CPS-SSA-IRS Exact Match File.

There is a data set that provides data relevant to accessing the differences in distributions of incomes between CPS nonrespondents and respondents. This data set is the CPS-SSA-IRS (SSA = Social Security Administration; IRS = Internal Revenue Service) Exact Match File (Aziz, Kilss, and Scheuren, 1978). The exact match file is based on a sample of 1978 CPS interviews with incomes obtained from SSA and IRS administration records. Thus, this file is a data set consisting of CPS respondents and nonrespondents with administrative income data always observed. By treating CPS nonrespondents' administrative income data as missing and applying specific methods for handling nonresponse, we do in fact obtain some evidence for the adequacy of these specific techniques for adjusting for nonresponse bias, although admittedly for administrative income rather than CPS reported income. Two papers doing this will be mentioned.

Herzog and Rubin (1982) compare the imputations from the CPS hot deck and an explicit two-stage linear/log-linear model; they also evaluate the utility of multiple imputation for obtaining proper inferences. This paper's objective, however, is to predict Social Security Benefits rather than total income, and so its results do not address the same kind of income nonresponse as studied in LSW.

A highly relevant paper, however, is Greenlees, Reece and Zieschang (1982), which also studies earned income. Not only does this paper use essentially the same selection model as LSW with the restriction $\theta = 0$ (i.e., income is log normal), but it also handles the truncation of income at $50,000 using maximum likelihood techniques. Interesting conclusions of this article are that (a) the model predicts

nonrespondent incomes rather well, (b) the true residuals in the log scale for the entire population, although not normal, are approximately symmetric, and (c) the CPS hot deck underestimates income by about 7%. These results lend modest, although mixed, support to the utility of the LSW selection model for CPS income data.

The results of applying the LSW techniques to the Exact Match File would certainly be of interest. Of particular importance, such an application could help to combat criticism based on the fact that the CPS data alone cannot be used to select which model for nonresponse is truly appropriate.

This suggestion, however, should certainly not be taken as indicative of a fatal flaw in LSW. LSW is an excellent paper. It is clearly written and demonstrates careful thought on important issues and fundamental understanding of the CPS and the Census Bureau's hot deck. Moreover, it describes extended statistical tools for handling the problem of nonresponse, and applies these tools to an important real world problem. LSW fits in very well with other important contributions on nonresponse.

REFERENCES

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.

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. Journal Royal Statistical Society B26: 211-252.

Cochran, W. G., and Rubin, D. B. (1973). Controlling bias in observational studies: A review. Sankhya - A, 35, 4: 417-446.

Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (1982). Imputation of missing values when the probability of reponse depends upon the variable being imputed. Journal of American Statistical Association.

Herzog, T. N., and Rubin, D. B. (1982). Using multiple imputations to handle nonresponse in sample surveys. Nonresponse in Sample Surveys: Theory and Bibliography. Report of the NAS Panel. Academic Press.

Little, R. J. A. (1982). Models for response in sample surveys. Journal American Statistical Association.

Rosenbaum, P. R., and Rubin, D. B. (1981). The central role of the propensity score in the analysis of observational studies for causal effects.

Rubin, D. B. (1976a). Multivariate matching methods that are equal percent bias reducing, I: Some examples. Biometrics, 32, 1: 109-120. Printer's correction note p. 955.

Rubin, D. B. (1976b). Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. Biometrics, 32, 1: 121-132. Printer's correction note p. 955.

Rubin, D. B. (1976c). Inference and missing data. Biometrika, 63, 3: 581-592.

Rubin, D. B. (1978a). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. With discussion and reply. The proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 20-34. Also in Imputation and editing of faulty or missing survey data. U. S. Department of Commerce, pp. 1-23.

Rubin, D. B. (1978b). Bayesian inference for causal effects: The role of randomization. The Annals of Statistics, 7, 1: 34-58.

Rubin, D. B. (1980a). Bias reduction using Mahalanobis' metric matching. Biometrics, 36, 2: 295-298. Printer's correction p. 296 (5,10) = 75%).

Rubin, D. B. (1980b). Handling Nonresponse in Sample Surveys by Multiple Imputations. U. S. Department of Commerce, Bureau of the Census Monograph.

DBR/db

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2358	2. GOVT ACCESSION NO. AD-A116 157	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) IMPUTING INCOME IN THE CPS: COMMENTS ON "WHAT DO WE KNOW ABOUT WAGES: THE IMPORTANCE OF NON-REPORTING AND CENSUS IMPUTATION" BY LILLARD, SMITH AND WELCH		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Donald B. Rubin		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 4 - Probability & Statistics
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE April 1982 13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

sample surveys, nonresponse, missing data, Box-Cox transformations, selection models

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Nonreporting of income in the Current Population Survey is an important problem affecting the many researchers using the data base. This paper discusses an approach to handling this problem proposed by Lillard, Smith and Welch, which applies selection models to a Box-Cox

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (continued)

transformation of the income variable. Topics considered here include: the inadequacy of single imputation and the desirability of multiple imputation, the importance of the distinction between ignorable and nonignorable nonresponse, the sensitivity of inference to assumptions unassailable by the data at hand, and the possibility of using the CPS-SSA-IRS Exact Match File to study such assumptions. The Lillard, Smith and Welch paper accompanied by this discussion is to appear in a book presenting the proceedings of the NBER Labor Cost Conference to be published by the University of Chicago Press.


DATE
ILME
8