MULTIPLE IMPUTATION OF INDUSTRY AND OCCUPATION CODES FOR PUBLIC-USE FILES[1]

Nathaniel Schenker, Donald J. Treiman, Lynn Weidman[2]
Nathaniel Schenker, UCLA School of Public Health, Los Angeles, CA 90024

1. INTRODUCTION

For each decennial census in the United States, the responses concerning employment are classified into industry and occupation categories. The industry and occupation coding schemes change somewhat each decade to allow an accurate representation of census employment information. Major changes were made for the 1980 census, especially for occupation classifications. The alterations in the coding schemes create difficulties in analyzing employment data from public-use samples over time, since the industry and occupation codes are not comparable across decades; for example, less than one-third of the 1970 occupation categories have an exact match in the 1980 classification (Treiman, Bielby, and Cheng 1988). Such cross-temporal analyses are important, as they yield information on topics such as job mobility and the effects of affirmative-action programs, which are of interest to both social scientists and government policy makers.

The problem of intertemporal comparability of industry and occupation classifications was studied recently by the Subcommittee on Comparability of Occupation Measurement (SCOM), which was sponsored jointly by the Social Science Research Council and the Bureau of the Census. Recent discussions of the problem are given in SCOM (1983) and Treiman, Bielby, and Cheng (1988). The SCOM (1983) suggested two possible methods for achieving comparability of industry and occupation codes in public-use samples from different censuses. The first is to assign codes using the 1980 coding scheme to public-use samples from previous censuses by directly coding the verbal responses for the people in the samples. This would be a very accurate method, but it would be prohibitively expensive for the 1960 and 1970 public-use samples, because the files released for public use do not include the verbal responses and these responses would be very costly to retrieve. The second method is to directly assign codes using the 1980 coding scheme to smaller samples from previous censuses, use these samples to estimate models predicting the 1980 codes from the old codes and covariates, and then use the models to impute 1980 codes for public-use samples from the previous censuses.

Imputation is a standard technique for handling missing items in surveys. It is especially well-suited to public-use data for two reasons: (1) the resulting completed data set can be analyzed using standard complete-data methods of analysis; and (2) the imputations can be created at the data production stage, when more is usually known by the data collector about the reasons for nonresponse than later by the typical user of the data set. A problem with imputing a single value for each missing item, however, is that when standard complete-data methods of analysis are applied to the completed data set, the uncertainty due to using imputed rather than true values is ignored, yielding inferences that are too sharp.

To allow the assessment of uncertainty associated with the imputation of 1980 industry and occupation codes for public-use samples from censuses prior to 1980, the SCOM (1983) suggested creating multiple imputations, a method proposed by Rubin (1978) and developed in detail in Rubin (1987). Multiple imputation replaces each missing datum with two or more values representing a distribution of likely values. The result is two or more completed data sets, each of which can be analyzed using the same complete-data method. These analyses can be combined to reflect both within-imputation variability and between-imputation variability as described in Section 2.

A project to multiply impute industry and occupation codes according to the 1980 coding scheme for two public-use samples from the 1970 census has just been completed with funding from the National Science Foundation and support from the Census Bureau (Treiman and Rubin 1983). A double-coded sample from 1970 (that is, a sample coded using both the 1970 and 1980 schemes) with 127,125 cases was used to estimate logit models predicting 1980 codes from the 1970 codes and covariates. The models were then used to impute five sets of 1980 codes for two one-percent public-use samples from 1970 with a combined total of 1.6 million cases. Rubin (1983), Rubin and Schenker (1987), and Treiman, Bielby, and Cheng (1988) describe the methods used for imputation in detail. In addition, a monograph describing the entire project is to be prepared.

This paper presents cross-temporal analyses of data from the 1970 and 1980 censuses. The data to be used from the 1970 census are one of the multiply-imputed public-use samples, with 794,100 cases, and the double-coded sample. Comparisons with 1980 will be made using a two-percent public-use sample from the 1980 census with 2.1 million cases. In all analyses, the industry and occupation codes from the 1980 coding scheme are used; these codes are known for the 1970 double-coded sample and the 1980 public-use sample and are imputed for the 1970 public-use sample.

The examples chosen for the paper are simple versions of analyses commonly of interest to researchers in the social sciences. The goals of the paper are (1) to compare multiple-imputation analyses with single-imputation analyses (that is, analyses using just one imputation), (2) to examine the utility of analyzing a large data set having imputed values (the public-use sample from 1970) versus analyzing a small data set having true values (the double-coded sample from 1970), and (3) to demonstrate and compare various methods of analyzing multiply-imputed data.

Recent theoretical and empirical work reported in Herzog and Rubin (1983), Rubin and Schenker (1986), Raghunathan (1987), and Rubin (1987) has shown that multiple imputation, even with just a few imputations per missing value, is superior to single imputation with regard to validity of interval estimates and significance levels. Comparisons between directly-assigned industry codes and multiply-imputed industry codes have been carried out by Rubin and Schenker (1987), Weld (1987), and Treiman, Bielby, and Cheng (1988), using the double-coded sample from 1970. These studies support the validity of analyses using multiply-imputed 1980 codes. The work of Treiman, Bielby, and Cheng (1988) suggests that inferences based on the 1970 public-use sample with multiply-imputed 1980 codes will usually be more precise than those based on the 1970 double-coded sample with true codes. We relate the properties of the cross-temporal analyses presented here to these previous studies.

Section 2 describes methods of analyzing a multiply-imputed data set. Section 3 presents analyses of changes in the sex composition of various occupations between 1970 and 1980. Section 4 presents comparisons between 1970 and 1980 of regressions of earnings on occupational status and sex for several industries in California. A summary is given in Section 5.

2. STATISTICAL INFERENCE FROM MULTIPLY-IMPUTED DATA

This section describes briefly the methods of analysis discussed in Rubin (1987, Chapter 3). Suppose that Q is a (vector-valued) quantity of interest and that with complete data, inferences for Q would be based on the statement that

$$(Q-\hat{Q}) \sim N(0,U), \qquad (2.1)$$

where $\hat{Q}$ is a statistic estimating Q, U is a statistic giving the dispersion matrix of $Q-\hat{Q}$, and N(0,U) denotes the multivariate normal distribution with mean 0 and dispersion matrix U.

In the presence of nonresponse, with m sets of imputations for the missing data, there are m sets of complete-data statistics, say $\hat{Q}_{*\ell}$ and $U_{*\ell}$, $\ell=1,\ldots,m$. In a multiple-imputation analysis, the m sets of complete-data statistics are combined as follows. Let

$$\bar{Q}_m = m^{-1}\sum_{\ell=1}^{m} \hat{Q}_{*\ell} \qquad (2.2)$$

be the average of the m complete-data estimates of Q,

$$\bar{U}_m = m^{-1}\sum_{\ell=1}^{m} U_{*\ell} \qquad (2.3)$$

be the average of the m complete-data dispersion matrices, and

$$B_m = (m-1)^{-1}\sum_{\ell=1}^{m}(\hat{Q}_{*\ell}-\bar{Q}_m)\ (\hat{Q}_{*\ell}-\bar{Q}_m)' \qquad (2.4)$$

be the between-imputation dispersion matrix of the m complete-data estimates of Q. The total variance of $(Q-\bar{Q}_m)$ is given by

$$T_m = \bar{U}_m+(1+m^{-1})B_m. \qquad (2.5)$$

When Q is scalar, interval estimates and significance levels are obtained using a t distribution with

$$\nu = (m-1)(1+r_m^{-1})^2 \qquad (2.6)$$

degrees of freedom, where

$$r_m = (1+m^{-1})B_m/\bar{U}_m \qquad (2.7)$$

is the ratio of the between-imputation component of variance to the within-imputation component. Thus, for example, a 100(1-α)% interval estimate of Q is

$$\bar{Q}_m \pm t_\nu(1-\alpha/2)T_m^{1/2},$$

where $t_\nu(1-\alpha/2)$ is the 1-α/2 quantile of the t distribution with ν degrees of freedom. The fraction of information about Q missing due to nonresponse is

$$\gamma_m = \frac{r_m+2/(\nu+3)}{r_m+1}. \qquad (2.8)$$

When Q is a k-dimensional (k>1) quantity of interest, there are several possible methods of finding the significance level of the null value Q_0 of Q. If m is large relative to k (e.g., $m\geq 5k$), the test statistic

$$D_m = (Q_0-\bar{Q}_m)'\,T_m^{-1}(Q_0-\bar{Q}_m)/k \qquad (2.9)$$

is referred to the F distribution with k and ν degrees of freedom; thus the significance level is

$$\Pr(F_{k,\nu}>D_m),$$

where $F_{k,\nu}$ is an F random variable with k and ν degrees of freedom. The denominator degrees of freedom ν is computed by (2.6) as used in the scalar case, but with (2.7) generalized to

$$r_m = (1+m^{-1})\mathrm{Trace}(B_m\bar{U}_m^{-1})/k. \qquad (2.10)$$

When m is modest relative to k, a test statistic that is better than D_m is

$$\tilde{D}_m = (1+r_m)^{-1}(Q_0-\bar{Q}_m)'\,\bar{U}_m^{-1}(Q_0-\bar{Q}_m)/k, \qquad (2.11)$$

which is referred to the F distribution with k and (k+1)ν/2 degrees of freedom.

Suppose a test of $Q=Q_0$ that is appropriate for complete data has been conducted using each of the m completed data sets resulting from multiple imputation. Let $p_{*1},\ldots,p_{*m}$ be the associated significance levels and let $d_{*\ell}$ be the value such that

$$\Pr(\chi_k^2>d_{*\ell}) = p_{*\ell},\ \ell=1,\ldots,m,$$

where χ_k^2 is a χ^2 random variable with k degrees of freedom; thus $d_{*1},\ldots,d_{*m}$ are the m complete-data χ^2 statistics associated with $\underset{\sim}{Q}_0$. A statistic asymptotically equivalent to D_m is

$$\hat{D}_m = \frac{\dfrac{\bar{d}_m}{k} - \dfrac{m-1}{m+1}r_m}{1+r_m}, \qquad (2.12)$$

where

$$\bar{d}_m = m^{-1}\sum_{\ell=1}^{m} d_{*\ell} \qquad (2.13)$$

is the average of the m complete-data χ^2 statistics; $\hat{D}_m$ is also referred to the F distribution with k and (k+1)ν/2 degrees of freedom. When only the m complete-data χ^2

statistics are available and r_m is not known, a method-of-moments estimator of r_m is

$$\hat{r}_m = \frac{(1+m^{-1})s_d^2}{2\bar{d}_m+[4\bar{d}_m^2-2ks_d^2]_+^{1/2}}, \qquad (2.14)$$

where

$$s_d^2 = (m-1)^{-1}\sum_{\ell=1}^{m}(d_{*\ell}-\bar{d}_m)^2 \qquad (2.15)$$

and $[\cdot]_+$ is equal to zero if the quantity inside the brackets is negative. Another test statistic, $\hat{D}_m^*$ ($\hat{D}_m$ in Rubin 1987), can be defined by replacing r_m with $\hat{r}_m$ in (2.12); this statistic is referred to the F distribution with k and $(1+k^{-1})\hat{\nu}/2$ degrees of freedom, where $\hat{\nu}$ is computed by replacing r_m with $\hat{r}_m$ in (2.6).

3. ANALYSES OF CHANGES IN THE SEX COMPOSITION OF OCCUPATIONS BETWEEN 1970 AND 1980

3.1 The Inference Problem

Suppose it is desired to estimate the change between 1970 and 1980 in the sex composition of an occupation defined according to the 1980 coding scheme. Let p_{70} and p_{80} denote the proportions of people in the occupation that are female for the 1970 and 1980 censuses, respectively; then $Q=p_{80}-p_{70}$. Given a sample from the 1970 census and a sample from the 1980 census, both having 1980 occupation codes for every case, Q is estimated by

$$\hat{Q} = \hat{p}_{80} - \hat{p}_{70}, \qquad (3.1)$$

where $\hat{p}_{70}$ and $\hat{p}_{80}$ are the sample proportions. The estimated variance of $\hat{Q}$ is

$$U = \hat{p}_{70}(1-\hat{p}_{70})/n_{70} + \hat{p}_{80}(1-\hat{p}_{80})/n_{80}, \qquad (3.2)$$

where n_{70} and n_{80} are the numbers of sample cases in the occupation. If n_{70} and n_{80} are not small, the complete-data inference for Q is given by the scalar version of (2.1), using (3.1) and (3.2).

3.2 Types of Analysis Examined

Three analyses of changes in sex composition are examined for each of the twelve occupations listed in Table 1. All three analyses use the 1980 public-use data; the differences in the analyses lie in which data from 1970 are used and how they are used. The numbers of sample cases in the twelve occupations are also given in Table 1.

Since the 1970 double-coded sample has true 1980 codes assigned to its cases, it is valid to perform the standard complete-data analysis given by (2.1), (3.1), and (3.2) using this sample and the 1980 public-use sample. A question of interest is whether inferences drawn in this way are more or less precise than inferences drawn from a multiple-imputation analysis involving the 1970 public-use sample.

As described in Section 1, five sets of 1980 codes were imputed onto the 1970 public-use sample. Thus, for each occupation, there are five values of $\hat{Q}$ and U as defined in (3.1) and (3.2) corresponding to the five values of $\hat{p}_{70}$ and n_{70}; note that p_{80} and n_{80} do not change across imputations because the true codes are known for the 1980 public-use sample. With the five sets of statistics, $\hat{Q}_{*\ell}$ and $U_{*\ell}$, $\ell=1,\ldots,5$, equations (2.2)-(2.8) can be used to perform a valid multiple-imputation analysis (with m=5).

It is common procedure to impute just one value for each missing item in a data set. As pointed out in Section 1, the application of standard complete-data methods of analysis to such a data set underestimates variability. To examine the magnitude of this underestimation, analyses involving the 1970 public-use sample but using only the first imputation will be presented. These analyses treat the one set of imputations as the true values and thus base inferences on the statement that

$$Q-\hat{Q}_{*1} \sim N(0,U_{*1}). \qquad (3.3)$$

3.3 Point Estimates

Table 2 displays point estimates of changes in the sex composition between 1970 and 1980 of the occupations defined in Table 1. Note that for the 1970 public-use sample, the estimates computed from just the first imputation ($\hat{Q}_{*1}$) are similar in value to the multiple-imputation estimates ($\bar{Q}_5$). Both estimates are unbiased (assuming the imputation model is unbiased) but $\bar{Q}_5$ is somewhat more efficient than $\hat{Q}_{*1}$ (Rubin 1987, Section 4.1). The estimates computed using the 1970 double-coded sample differ from $\hat{Q}_{*1}$ and $\bar{Q}_5$ more than $\hat{Q}_{*1}$ and $\bar{Q}_5$ differ from each other. This is due to sampling variability, as the double-coded sample is only one-sixth the size of the public-use sample.

3.4 Comparisons of Validity and Precision

The main issues to be investigated here are the underestimation of variability when a single-imputation analysis is performed rather than the valid multiple-imputation analysis and the precision of inferences involving the 1970 double-coded sample relative to inferences involving the multiply-imputed 1970 public-use sample. Table 3 presents measures of variability computed from the three types of analysis.

Multiple imputation versus single imputation

Column (2) of Table 3 displays the standard errors $U_{*1}^{1/2}$ for the single-imputation analysis based on (3.3). Column (5) shows the standard errors $T_5^{1/2}$ for the t-based multiple-imputation analysis described in Section 2; the degrees of freedom ν for the t distribution are given in column (6). As mentioned in Section 1, previous work has shown that single-imputation analyses tend to yield inferences that are too sharp because the between-imputation variability is then ignored. Single-imputation inferences are further from being valid when the fraction of information missing due to nonresponse is high because a large proportion of the total variability is then from the between-imputation component. Column (9) shows the ratios of the single-imputation standard errors to the multiple-imputation standard errors. In six out of twelve cases, this ratio is less than 0.75; this indicates, for example, that in these cases single-imputation interval estimates will be less than three-fourths as wide as they should be. Note that the ratio tends to be lower when the fraction of information missing due to nonresponse (column (7)) is higher, and that the ratio is nearly one when the fraction of missing information is very small (occupation 808).

1970 double-coded sample versus multiply-imputed 1970 public-use sample

The standard errors $U^{1/2}$ for the analysis involving the 1970 double-coded sample based on (2.1), (3.1), and (3.2) are displayed in column (1) of Table 3. In comparing this analysis with the multiple-imputation analysis involving the 1970 public-use sample, the issue is whether the loss in precision due to the smaller size of the double-coded sample is greater than the loss of precision due to having imputed rather than true codes for the public-use sample. Column (8) shows the ratio of $U^{1/2}$ to $T_5^{1/2}$. Although both analyses are valid, the multiple-imputation analysis involving the public-use sample tends to yield more precise inferences than the analysis involving the double-coded sample. In seven out of twelve cases the ratio in column (8) is greater than 1.5. Only for occupation 263 is the standard error using the double-coded sample smaller than the standard error using the multiply-imputed public-use sample.

It appears that analyses involving the 1970 public-use sample with multiply-imputed codes usually yield more precise inferences than analyses involving the 1970 double-coded sample; this was predicted by Treiman, Bielby, and Cheng (1988). Another related advantage of the multiply-imputed public-use sample is that the larger sample size makes more detailed analyses possible. For instance, although the analyses presented here are for the entire United States, the double-coded sample has only 25 observations for occupation 583 whereas the public-use sample has over 150 observations. If finer geographical detail were desired, many analyses would be impossible using the double-coded sample but still possible using the public-use sample.

3.5 A Comment on the Fraction of Missing Information

Note that although the simple nonresponse rate for the 1970 public-use sample is 100% (all 1980 industry and occupation codes are imputed), the fractions of information missing due to nonresponse in column (7) of Table 3 range from 2% to 78%. The fraction of missing information, γ_m, estimates the relative difference between the Fisher information about Q in the multiple-imputation t distribution (Section 2) and the Fisher information about Q that would exist with complete response (Rubin 1987, Section 3.3). It is based on the size of the between-imputation variability relative to the within-imputation variability. Thus it takes into account the precision of the imputation model (which depends on the predictive power of covariates used as well as the simple nonresponse rate) and the specific inference problem being considered.

4. ANALYSES OF CHANGES IN REGRESSIONS OF EARNINGS ON OCCUPATIONAL STATUS AND SEX BETWEEN 1970 AND 1980

4.1 The Inference Problem

Suppose it is desired to test whether the slope coefficients of the linear regression of earnings on occupational status (as measured by Duncan's socio-economic index, updated to 1980 by Stevens and Cho) and sex (0=male, 1=female) have changed between 1970 and 1980, where occupations are defined according to the 1980 coding scheme. Let β_{70} and β_{80} denote the vectors of slope coefficients for the 1970 and 1980 censuses. Then the goal is to test whether $(\beta_{80}-\beta_{70})$ is equal to zero; thus $Q=\beta_{80}-\beta_{70}$ and $Q_0=0$ in the notation of Section 2. Given samples from the 1970 and 1980 censuses, both containing 1980 codes, Q is estimated by

$$\hat{Q} = b_{80} - b_{70}, \qquad (4.1)$$

where b_{70} and b_{80} are the least-squares estimates of β_{70} and β_{80} obtained from the samples. If V_{70} and V_{80} are the estimated dispersion matrices of b_{70} and b_{80} (obtained by multiplying the inverse of the matrix of sums of squares and cross products by the error mean square), then the estimated dispersion matrix of $\hat{Q}$ is

$$U = V_{70} + V_{80}. \qquad (4.2)$$

If the sample sizes are not small, then complete-data inferences for Q are based on (2.1), (4.1), and (4.2). For example, a test of whether Q=0 is carried out by referring the test statistic

$$d = \hat{Q}' U^{-1} \hat{Q} \qquad (4.3)$$

to the χ^2 distribution with two degrees of freedom.

4.2 Types of Analysis Examined

With five sets of 1980 industry and occupation codes imputed for the 1970 public-use sample, there are five values of $\hat{Q}$ and U as defined in (4.1) and (4.2) corresponding to the five values of b_{70} and V_{70}; note that b_{80} and V_{80} do not change across imputations because the true codes are known for the 1980 public-use sample. The five sets of estimates and dispersion matrices, $\hat{Q}_{*\ell}$ and $U_{*\ell}$, $\ell=1,\ldots,5$, can be used to compute the multiple-imputation test statistics D_5 and $\tilde{D}_5$ (equations (2.9) and (2.11)) as described in Section 2 (with m=5). Alternatively, the complete-data test statistic (4.3) can be computed five times, yielding

$$d_{*\ell} = \hat{Q}'_{*\ell} U^{-1}_{*\ell} \hat{Q}_{*\ell}, \quad \ell=1,\ldots,5;$$

these five χ^2 statistics may be combined as described in equations (2.12)-(2.15) to compute $\hat{D}_5$ and $\hat{D}_5^*$.

The comparison of regression coefficients is carried out for data from California, with separate comparisons made within each of the seven industries listed in Table 4. (Note: In the regression calculations for the 1980 public-use sample, the earnings variable was reexpressed in 1970 dollars using the Consumer Price Index to make the regression results for 1970 and 1980 numerically comparable.) Table 4 also lists the number of cases in the seven industries for the 1970 and 1980 public-use samples.

Research on significance testing has shown that multiple-imputation tests have actual levels (rejection probabilities under the null hypothesis) that are much closer to the nominal levels than is the case for single-imputation tests (Raghunathan 1987; Rubin 1987, Section 4.8). For the testing problem considered here, however, it is impossible to demonstrate the invalidity of single-imputation tests because the true values of the regression coefficients for each industry are unknown and thus it is unknown whether the null hypothesis is true. Similarly, it is difficult to compare the power of tests using the 1970 public-use sample with tests using the 1970 double-coded sample because even if the null hypothesis is not true, the true alternative is unknown. The focus here will therefore be restricted to demonstrating and comparing the four methods of testing using multiply-imputed data.

4.3 Comparison of Multiple-Imputation Tests

Columns (1)-(4) of Table 5 display the test statistics D_5, $\tilde{D}_5$, $\hat{D}_5$, and $\hat{D}_5^*$ for the seven industries defined in Table 4, with the descriptive significance levels (p-values) given in parentheses. All four test statistics have been referred to F distributions with k=2 degrees of freedom for the numerator; the denominator degrees of freedom (see Section 2) are ν for D_5, 1.5ν for $\tilde{D}_5$ and $\hat{D}_5$, and .75$\hat{\nu}$ for $\hat{D}_5^*$, where the values of ν and $\hat{\nu}$ are given in columns (3) and (5) of Table 6.

Comparison of columns (2) and (3) of Table 5 shows that the asymptotically equivalent statistics $\tilde{D}_5$ and $\hat{D}_5$ are very similar for the seven industries considered here. A general advantage of using $\hat{D}_m$ rather than $\tilde{D}_m$ is that the computation of $\hat{D}_m$ requires m scalar χ^2 statistics rather than m sets of k-vector estimates and kxk estimated dispersion matrices. This is useful in practice, since complete-data methods of testing a hypothesis about a multidimensional Q often produce a single test statistic rather than an estimate and dispersion matrix. In such cases, however, the value of r_m is not readily available. An approximate solution is to replace r_m with $\hat{r}_m$ (equation (2.14)), yielding $\hat{D}_m^*$. Column (4) of Table 5 displays the values of $\hat{D}_5^*$ for the seven industries. The largest differences between $\hat{D}_5^*$ and $\hat{D}_5$ occur when $\hat{r}_5$ is very different from r_5 (see columns (1) and (4) of Table 6). Note that the value of $\hat{D}_5^*$ for industry 351 is negative, which is inappropriate for a test statistic that uses an F reference distribution. Rubin (1988) discusses current research on improving the test statistic $\hat{D}_m^*$.

For each of the seven industries, the four multiple-imputation tests yield qualitatively the same results. The widest ranges in p-values occur for industires 351, 901, and 902, for which the fractions of missing information are 28%, 60% and 52%, respectively (column (2) of Table 6). The four test statistics have very similar values for industry 440, which only has 4% missing information.

5. SUMMARY OF RESULTS

Two cross-temporal inference problems were discussed in this paper: estimating changes in the sex composition of occupations between 1970 and 1980, and testing whether the slope coefficients of regressions of earnings on occupational status and sex have changed between 1970 and 1980. Three types of analysis were examined, each of which used the same 1980 public-use data but different data for 1970. The three 1970 data sets were the double-coded sample with true 1980 codes, the public-use sample with five sets of imputations of the 1980 codes, and the public-use sample with just one set of imputations.

For the estimation problem, the invalid single-imputation analyses always produced standard errors that were smaller than those produced by the valid multiple-imputation analyses; this corroborates previous theoretical and Monte Carlo results showing that single-imputation analyses yield interval estimates that are too short and have lower than nominal coverage probabilities (Herzog and Rubin 1983; Rubin and Schenker 1986, 1987; Rubin 1987). The multiple-imputation analyses outperformed the valid complete-data analyses involving the 1970 double-coded sample with respect to precision of inferences. Interval estimates from the multiple-imputation analyses were usually narrower than those from the analyses involving the double-coded sample. These results confirm the research of Treiman, Bielby, and Cheng (1988).

Four multiple-imputation methods of testing were applied in the regression problem. The results for the four methods were qualitatively similar. The asymptotically equivalent test statistics $\tilde{D}_m$ and $\hat{D}_m$ behaved almost identically. The statistic $\hat{D}_m^*$ did not always yield an accurate approximation to $\hat{D}_m$. In addition, one negative value of $\hat{D}_m^*$ occurred. Current research on improvements to $\hat{D}_m^*$ are described in Rubin (1988).

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

[2] Nathaniel Schenker is in the Division of Biostatistics, School of Public Health and Donald Treiman is in the Department of Sociology, both at UCLA. Lynn Weidman is in the Statistical Research Division, U.S. Bureau of the Census.

REFERENCES

Herzog, T.N., and Rubin, D.B. (1983), "Using Multiple Imputations to Handle Nonresponse in Surveys," in *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*, W.G. Madow, I. Olkin, and D.B. Rubin (eds.), New York: Academic Press, pp. 209-245.

Raghunathan, T.E. (1987), *Large Sample Significance Levels from Multiply-Imputed Data*, Ph.D. Thesis, Department of Statistics, Harvard University.

Rubin, D.B. (1978), "Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34.

Rubin, D.B. (1983), "Progress Report on Project for Multiple Imputation of 1980 Codes," distributed to the Bureau of the Census, the National Science Foundation, and the Social Science Research Council.

Rubin, D.B., (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D.B. (1988), "An Overview of Multiple Imputation," *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Rubin, D.B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random samples with Ignorable Nonresponse.," *Journal of the American Statistical Association*, 81, 366-374.

Rubin, D.B., and Schenker, N. (1987), "Interval Estimation from Multiply-Imputed Data: A Case Study Using Census Agriculture Industry Codes," *Journal of Official Statistics*, 3, 375-387.

Subcommittee on Comparability of Occupation Measurement (1983), "Alternative Methods for Effecting the Comparability of Occupation Measurement over Time," report to the Social Science Research Council and the Bureau of the Census.

Treiman, D.J., Bielby, W., and Cheng, M. (1988), "Evaluating a Multiple-Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard," to appear in *Sociological Methodology*, 18.

Treiman, D.J., and Rubin, D.B. (1983), "Multiple Imputation of Categorical Data to Achieve Calibrated Public-Use Samples," proposal to the National Science Foundation.

Weld, L.H. (1987), *Significance Levels from Public-Use Data with Multiply-Imputed Industry Codes*, Ph.D. Thesis, Department of Statistics, Harvard University.

TABLE 1

OCCUPATIONS ANALYZED FOR CHANGES IN SEX COMPOSITION
BETWEEN 1970 AND 1980

1980 Code	Occupation	Number of Cases in Occupation		
		1970 Double-coded Sample	1970 Public-Use Sample *	1980 Public-Use Sample
067	Statisticians	49	225	562
084	Physicians	482	2872	8645
095	Registered nurses	1162	8345	25695
263	Sales workers, motor vehicles and boats	345	2571	5632
375	Insurance adjusters, examiners, and investigators	168	1185	3332
418	Police and detectives, public service	515	3382	8401
484	Nursery workers	33	186	693
583	Paperhangers	25	152	315
686	Butchers and meat cutters	473	3060	5952
704	Lathe and turning machine operators	224	1284	2370
808	Bus drivers	408	2579	7844
889	Laborers, except construction	1323	8652	27992

*For the 1970 public-use sample, the number of cases varies across imputations. The minimum number is given here.

TABLE 2

ESTIMATES (IN PERCENT) OF CHANGES IN THE SEX COMPOSITION OF OCCUPATIONS BETWEEN 1970 AND 1980

Occupation	1970 Double-Coded	1970 Public-Use Sample	
(1980 code)	Sample ($\hat{Q}$)	$\hat{Q}_{*1}$	$\bar{Q}_5$
067	14.95	5.38	4.93
084	5.34	4.06	3.54
095	-2.55	-1.45	-1.58
263	5.17	-0.13	0.69
375	30.45	30.50	30.32
418	3.53	2.16	2.02
484	-2.30	-8.66	-13.83
583	10.10	6.12	8.42
686	2.89	2.17	2.76
704	-3.98	1.09	-0.37
808	18.69	18.17	18.01
889	3.91	5.14	5.00

NOTE: Positive values indicate an increase in the percent female from 1970 to 1980.

TABLE 3

MEASURES OF VARIABILITY FOR THE ANALYSIS OF CHANGES IN THE (PERCENT) SEX COMPOSITION OF OCCUPATIONS BETWEEN 1970 AND 1980

Occupation	1970 Double-Coded	1970 Public-Use Sample						Ratio:	Ratio:
(1980 code)	Sample ($U^{1/2}$)	$U_{*1}^{1/2}$	$\bar{U}_5$	$1.2B_5$	$T_5^{1/2}$	ν	γ_5(%)	(1)/(5)	(2)/(5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
067	7.12	3.87	14.88	1.10	4.00	839	7	1.78	0.97
084	1.31	0.66	0.45	0.44	0.95	16	55	1.38	0.70
095	0.38	0.21	0.04	0.01	0.24	91	23	1.62	0.90
263	0.88	0.62	0.36	0.97	1.15	8	78	0.77	0.53
375	3.66	1.56	2.49	1.04	1.88	46	32	1.95	0.83
418	0.71	0.41	0.17	0.16	0.57	18	52	1.24	0.72
484	8.90	4.11	12.77	33.68	6.81	8	78	1.31	0.60
583	5.84	3.32	9.71	4.37	3.75	42	34	1.56	0.88
686	1.53	0.74	0.54	0.13	0.82	104	21	1.87	0.91
704	2.28	0.90	0.91	1.01	1.38	14	58	1.65	0.65
808	2.26	1.04	1.08	0.02	1.05	9427	2	2.15	0.99
889	1.03	0.44	0.19	0.47	0.81	8	76	1.27	0.54

TABLE 4
INDUSTRIES ANALYZED FOR CHANGES IN REGRESSION COEFFICIENTS IN CALIFORNIA BETWEEN 1970 AND 1980

1980 Code	Industry	Number of Cases in Industry	
		1970 Public-Use Sample *	1980 Public-Use Sample
351	Manufacturing: motor vehicles and motor vehicle equipment	407	1280
440	Radio and television broadcasting	142	530
591	Retail trade: department stores	1762	4457
700	Banking	1163	4134
850	Colleges and universities	1715	4807
901	General government	343	2932
922	Administration of human resources programs	286	931

* For the 1970 public-use sample, the number of cases varies across imputations. The minimum number is given here.

TABLE 5

MULTIPLE-IMPUTATION TEST STATISTICS (P-VALUES (%) IN PARENTHESES) FOR CHANGES IN REGRESSION COEFFICIENTS IN CALIFORNIA BETWEEN 1970 AND 1980

Industry	D_5	$\tilde{D}_5$	$\hat{D}_5$	$\hat{\hat{D}}_5$
(1980 code)	(1)	(2)	(3)	(4)
351	0.14 (87)	0.16 (85)	0.17 (84)	-0.00 (100)
440	1.14 (33)	1.1 (33)	1.11 (33)	1.17 (31)
591	6.99 (.11)	6.80 (12)	6.80 (.12)	7.18 (.082)
700	6.73 (.16)	5.81 (.35)	5.87 (.33)	6.58 (.15)
850	1.43 (24)	1.33 (.26)	1.34 (26)	1.47 (23)
901	0.32 (73)	0.28 (76)	0.26 (77)	0.94 (40)
922	1.71 (21)	1.30 (29)	1.27 (30)	1.91 (16)

TABLE 6

RATIOS OF VARIABILITY, FRACTIONS OF MISSING INFORMATION, AND DEGREES OF FREEDOM FOR MULTIPLE-IMPUTATION ANALYSES OF CHANGES IN REGRESSION COEFFICENTS IN CALIFORNIA BETWEEN 1970 AND 1980

Industry	r_5(%)	γ_5(%)	ν	$\hat{r}_5$(%)	$\hat{\nu}$
(1980 code)	(1)	(2)	(3)	(4)	(5)
351	34.5	28	61	69.2	24
440	4.0	4	2759	0.5	148934
591	12.2	11	337	6.8	995
700	20.4	18	139	8.5	655
850	8.7	8	622	2.0	10619
901	121.7	60	13	28.0	84
922	90.1	52	18	43.2	44