STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: Census/SRD/89/03

Final Report
INDUSTRY AND OCCUPATION IMPUTATION

by

Lynn Weidman
Statistical Research Division
Bureau of the Census
Room 3134, F.O.B. #4
Washington, D.C. 20233 U.S.A.

Recommended by: Lawrence Ernst

Report Completed: April 20, 1989

Report Issued: April 20, 1989

FINAL REPORT ON PROJECT #85-2
INDUSTRY AND OCCUPATION IMPUTATION
Project Manager: Lynn Weidman

Preface

The Industry and Occupation Imputation Project was undertaken as a joint effort of the Bureau of the Census and the Social Science Research Council (SSRC), with the goal of producing 1970 census public use files containing industry and occupation codes that are directly comparable with the corresponding codes on the 1980 public use files. This would allow researchers to analyze changes in job-related characteristics that took place during the decade of the 1970's. This report is intended as documentation of the statistical techniques used in the project, and summarizes the software development, computational and data-manipulation tasks performed to implement them and produce the necessary files. A detailed history of the project will be produced by other participants.

The Census Bureau's participation in this project has been under the direction of Thomas Scopp, Chief, Labor Force Statistics Branch, Housing and Household Economic Statistics Division (formerly in Population Division). Statistical Research Division (SRD) participants included Lynn Weidman, Nathaniel Schenker and Bradley Schultz (one year). Other major Census Bureau contributors were John Priebe (Labor Force Statistics Branch) and Len Marshall (Population Division). Primary outside advisors providing statistical and labor force expertise were Clifford Clogg (Pennsylvania State University), Donald Rubin (Harvard) and Donald Treiman (UCLA).

The rest of the preface summarizes the role of SRD in the project. SRD's participation originated in the fall of 1982 when Population Division requested the involvement of a mathematical statistician to work with the outside advisors on developing and applying statistical techniques to estimate a large number of logit models for use in multiple imputation of industry and occupation codes onto census public use files. Lynn Weidman began working on the project at that time and continued as primary SRD participant, with varying percentages of involvement until the production of the final imputed files was completed in December 1988. Bradley Schultz worked full time on the project during his 14 month tenure at the Census Bureau. Nathaniel Schenker began working on the project as a graduate student of Don Rubin's, and continued on it part-time throughout his three year Census Bureau tenure and at UCLA until its completion.

SRD took on the role of actually producing the multiply-imputed records, as well as participating in development and evaluation of proposed statistical methods. As a result, major responsibilities of SRD participants have included implementing the statistical methodology of this project via design and development of all software used, performing all evaluative and production computer runs of the software used, determining the industries and occupations that require model estimation, and evaluating the results from some of the statistical procedures employed or proposed.

1. Executive Summary

There is a great deal of interest among social science researchers in comparing industry and occupation (I&O) data over time. The occupation classification system used by the Census Bureau, however, was considerably revised between the 1970 and 1980 censuses, making occupation comparability between 1980 and previous censuses virtually impossible. The same type of problem exists for industry, but to a much lesser extent. In order to establish a basis for I&O comparability, a benchmark set of records coded under both the old and revised classification systems was needed. It was determined that double-coding millions of records by hand would be too expensive. Instead, records from 1970 public use samples had multiple imputations put on them. Multiple imputation, as opposed to single imputation, enables analyses of the imputed data sets to reflect variance in classification due to the imputation procedure.

A series of dichotomous logistic regressions is used to represent the probability P_{ij} that a person who was classified as being in occupation i under the 1970 classification system would be classified as being in occupation j under the 1980 system. These probabilities are functions of individual demographic and socio-economic characterisitics and unknown parameters. For a given 1970 occupation (industry) the parameters of the models are estimated using the appropriate records from a 127,125 record file that has been double-coded under the two classification systems.

Several sets of parameter values are randomly generated from the asymptotic normal posterior distribution of the estimated parameters for each model. Five of these sets are then selected for use in imputation according to a sampling/importance resampling mechanism. Each record from two 1% 1970 census county group public use samples has five occupations (industries) imputed using the P_{ij}'s calculated from the appropriate 5 sets of parameters.

An outline of the major steps in this project follows.

A. Write software to perform each of the three major statistical tasks.
   1. Estimation of the parameters of binomial logit models using simple Bayesian priors.
   2. Generation of parameter sets from the normal posterior distribution of parameter estimates and selection by the sampling/importance resampling procedure of five of them for further use.
   3. Multiple imputation of industries and occupations onto public use files using the selected parameter sets.

B. Determine which 1970 industries require modeling. The remaining industries have unique 1980 codes.

C. For each industry requiring modeling, construct a file of records having that industry code from the double-coded file. Construct similar files for each major group.

D. Estimate and select parameters for each model using A.1 and A.2.

E. Multiply impute five industry codes onto records from the double-coded file using A.3. Evaluate the quality of this imputation.

F. Perform B-E for occupations.

G. Using selected 1970 census public use files, construct separate files of person records for each industry and major group code requiring imputation, and put the rest of the person records in a single file. Do the same for occupation codes.

H. Assign five imputes to each of these eligible records, one file at a time, using A.3.

I. Recombine these records into new public use files that are identical to the original except that each eligible record now has five imputed 1980 industry and five imputed 1980 occupation codes.

J. Copy these files to tape and give to DUSD for public release.

2. Selecting 1970-1980 Code Combinations to be Modeled

A sample of 127,125 records from the microfilm of original 1970 census long form questionnaires was randomly selected for use in this project. All of these records contained the original I&O codes assigned in the 1970 census and were also assigned I&O codes according to the 1980 classification scheme. This set of "double-coded" records gives us a database which can be used to model the relationships between the 1970 and 1980 coding schemes. The question of interest to us is how these relationships can be used to impute 1980 codes to a much larger set of 1970 public use records. Comparisons between I&O distributions in 1970 and 1980 can then be carried out using these imputed records and 1980 public use files.

2.1 Use of Distribution of Codes

Initially, a determination of which 1970 industry codes required modeling had to be made. For each 1970 industry the observed distribution of assigned 1980 industry codes was calculated. All industries that mapped into at least two 1980 industries, with at least two records each, required modeling. There are 180 of these industries, and for each of them a separate file of records was constructed for use in model estimation and imputation of the double-coded file. For each of the remaining 34 industries (called 100% industries), its single 1980 code would always be imputed.

This same procedure was used for occupations except for one added feature. Some occupations with less than 50 records were aggregated for modeling if they had common 1980 codes. Altogether there are 249 single occupations requiring models, 55 occupations combined into 26 aggregates that require modeling, 118 100% occupations and 4 equal probability (see section 2.3) occupations. (See Tables 1 and 2 for detailed breakdowns by industry and occupation.) In addition, there are 1970 records that were imputed major industry and/or occupation group codes in the 1970 processing rather than individual codes, usually because there were no written entries to code clerically. These 14 major industry and 12 major occupation groups also required modeling.

2.2 Records with Unique Code Combinations

The double-coded sample includes only some of the records that were assigned 1970 occupation codes in census processing and does not represent all 1970-1980 pairs that would occur in the full set of records. However, we are not modeling any of these possible pairs that did not occur in the double-coded sample. Also, some records were excluded from the modeling procedure as being "unique," as described in the next paragraph.

Following selection of the 127,125 record sample, a second 1970 I&O coding was performed and differences between the two 1970 codings resolved, so that each record of the double-coded file had an original 1970 code, a "corrected" 1970 code and a 1980 code. A lot of original 1970-corrected 1970-1980 combinations occur only once in the double-coded sample. For any particular combination this might be the result of an uncommon error, such as assigning an original code that was unrelated to the correct code or a keypunch error made when transcribing the codes, and could be the only such error in the census. For a more exact imputation, records with these unique combinations should be modeled. However, they would require the use of a substantially larger number of models (smaller than the number of such records, but one for each 1980 code not already being modeled) and have almost no affect on most substantive analyses. For these reasons we have chosen, with two exceptions, not to use these records. If a particular combination occurs more than once in the double-coded sample, then we know it was not a unique error that occured. The 1970-1980 pair from this combination probably exists with reasonable probability in the census, and we include all such records in model estimation.

There are two exceptions to non-use of records with unique combinations for individual 1970 codes. The first is the aggregated occupations, where a single distribution of combinations is computed for all records in the aggregate before unique combinations are identified. The second is four occupations that only have unique combinations. Since all the cases can't be excluded for these industries, each of the 1980 codes is treated as having probability equal to the proportion of these records with that 1980 code.

3. Selecting a Class of Probability Models

For each group of records with a given 1970 occupation code, there is a corresponding set of 1980 occupation codes, as described in the previous section. The probability of a person having any of these 1980 codes can be represented as a function of selected demographic characteristics. Our approach was to estimate these probabilistic relationships between the coding schemes using logit models. A multinomial logit could have been used to model the probabilities for all 1980 codes simultaneously. However, there are certain drawbacks to this approach:

a) it requires simultaneous estimation of a large number of parameters;
b) the accuracy of parameter estimates depends on the amount of data available for each possible 1980 code, so the quality of estimates for more populous codes is affected by lack of data for the less populous codes;
c) little software is available for estimating multinomial logit models, and it is not easily modifiable for our needs.

Due to these drawbacks we decided to model a series of bivariate logit models. First we modeled the most populous code against all the rest, then the second most populous against all less populous codes, etc., ending with a model for the two least populous codes.

Example. Consider 1970 industry 859 (Libraries) and the 1980 industry codes into which it mapped on the double-coded file.

1980 code	852	850	841	842
# cases	189	8	3	2

A series of three models was used:

1) 852 vs 850,841,842 (189 cases vs 13)
2) 850 vs 841,842 (8 cases vs 5)
3) 841 vs 842 (3 cases vs 2)

3.1 Binomial Logit Models

Let $\underline{x}_i$ be a column vector of (dummy) variables for person i and $\underline{\beta}$ a corresponding vector of unknown parameters. A logit model for the probability that a binomial (0,1) variable Y can take on the value 1 is

$$\text{Probability } (Y_i=1) = \exp(\underline{x}_i'\underline{\beta})/[1 + \exp(\underline{x}_i'\underline{\beta})]. \qquad (3.1.1)$$

In our situation the dummy variables represent categories of demographic predictors and $Y_i=1$ if person i is in the most populous occupation being modeled and 0 otherwise.

3.2 Selection of Predictor Variables

When estimating a model for the purpose of prediction, common practice is to select the best-fitting model with the smallest number of terms from a potential set of variables and their interactions. In our situation we are interested in both prediction (imputation) and the analyses that will be made with the resulting data. If a variable is excluded from a model, then no variability in the estimated parameters and imputed values due to this variable will be reflected. As a result, variables used in analyses but not in the model will show no effect on the variability between imputations, although they may affect it quite differently. Thus, it is important to include substantively important variables in the models. On the other hand, it ~~is~~ not possible to include all such variables, so some selectivity was required.

The approach used to determine appropriate variables was to ask knowledgeable social scientists which variables were most important to and most often used by analysts of I&O data. Members of the Social Science Research Councils's Subcommittee on the Comparability of Occupational Measurement were polled on what variables should be used. The responses were combined and refined by members of the project team to produce the final list given in Table 3. All variables were treated as categorical using the category definitions listed in Table 4.

4. Model Estimation

The vector $\underline{\beta}$ of unknown parameters in the model (3.1) is estimated by maximum likelihood. If there are a total of N observations being used to estimate the model, the likelihood function is

$$\prod_{i=1}^{N} \left[\frac{\exp(\underline{x}_i'\underline{\beta})}{1+\exp(\underline{x}_i'\underline{\beta})}\right]^{y_i} \left[\frac{1}{1+\exp(\underline{x}_i'\underline{\beta})}\right]^{1-y_i} \qquad (4.1)$$

All the predictor variables we are using are categorical. Consider the cross-classification of the categories of these predictor variables with each combination (possible value of $\underline{x}_i$) defining one cell of a table. Every person with a particular $\underline{x}_i$ belongs to the same cell and makes the same contribution to the likelihood function. Letting C=number of cells, n_{j1}=number of persons in cell j with Y_i=1 and n_{j0}=number of persons in cell j with Y_i=0, the likelihood function can be rewritten as

$$\prod_{j=1}^{C} [\exp(\underline{x}_j'\underline{\beta})]^{n_{j1}} [1 + \exp(\underline{x}_j'\underline{\beta})]^{-n_{j0}-n_{j1}}$$

4.1 The Sparse Data Problem

A sufficient condition for the existence of maximum likelihood estimators (mle's) for all elements of $\underline{\beta}$ is that $n_{j0}, n_{j1} > 0$ for all j. If some of the counts are 0, the existence of mle's for a specified model and table is difficult to determine. The industry models each have 2304 cells with the number of observations per model between 4 and 3500. In all cases there are many cells for which n_{j0} and/or n_{j1} are 0, and in the worst cases there are at least 2300 cells for which $n_{j0}+n_{j1}=0$. This situation also holds for occupation models which have even more cells. We needed to develop a procedure which would solve this estimability problem in a way that would allow us to automate the solution for use with a large number of parameter estimations.

4.2 Adding Prior Data to Ensure Estimability

One way to meet the sufficiency condition is to put a small amount of data into each cell of a table for both Y=1 and Y=0, i.e., give n_{j0} and n_{j1} some small non-zero value for all j. Some questions to consider in this approach are: (a) how much should be added to each n_{j0} and n_{j1}?; (b) should n_{j0} and n_{j1} get the same amounts?; (c) how does adding this data affect parameter estimates? Before answering these questions, we show that this method is equivalent to using a certain prior distribution in a Bayesian approach to estimating $\underline{\beta}$. First, for cell j, let $\ell_j = \underline{x}_j' \underline{\beta}_j$ and $\pi_j(\ell_j)$ be P(Y=1) as given by (3.11). Then the observations in cell j are binomial with

$$P(Y=1) = \pi_j(\ell_j) = \exp(\ell_j)/[1 + \exp(\ell_j)], \text{ where}$$

$$\ell_j = \underline{x}_j'\underline{\beta} = \ln[\pi_j(\ell_j)/(1-\pi_j(\ell_j))] = \text{logit } (\pi_j).$$

This cell's contribution to the overall likelihood function is $[\pi_j(\ell_j)]^{n_{j1}}[1-\pi_j(\ell_j)]^{n_{j0}}$, which has the beta distribution as the natural conjugate prior,

$$P(\pi_j) = [\pi_j(\ell_j)]^{\alpha_1}[1-\pi_j(\ell_j)]^{\alpha_0} \text{ with } \alpha_{j0},\alpha_{j1} > -1. \qquad (4.2.1)$$

Using this prior, the posterior distribution of $\pi_j(\ell_j)$ is proportional to

$$[\pi_j(\ell_j)]^{n_{j1}+\alpha_{j1}} [1-\pi_j (\ell_j)]^{n_{j0}+\alpha_{j0}} \qquad (4.2.2)$$

which is equivalent to adding α_{j1} successes and α_{j0} failures to the observed data in cell j.

Now, in order to describe the prior used in this project, define

- p = number of parameters in a model,
- C = number of cells in a table,
- $n_{\cdot 0}$ = number cases in table with Y=0,
- $n_{\cdot 1}$ = number cases in table with Y=1,
- s = $n_{\cdot 1}/(n_{\cdot 0}+n_{\cdot 1})$ = proportion of cases with Y=1.

Then the prior used is to add

$$\alpha_{ji} = \alpha_1 = sp/C \text{ to each } n_{j1}, \qquad (4.2.3a)$$

$$\alpha_{j0} = \alpha_0 = (1-s)p/C \text{ to each } n_{j0}. \qquad (4.2.3b)$$

This prior has the following properties:

it treats all cells equally;
it pulls the estimate of the constant parameter toward $\ln[s/(1-s)]$;
it pulls the other parameter estimates toward 0;
the total amount of prior data added is p;
for all tables with C cells and models with p parameters it is a noninformative prior.

4.3 Estimation Procedure

After the addition of prior data the likelihood function is

$$\prod_{j=1}^{C} [\exp(\underline{x}_j'\underline{\beta})]^{n_{j1}+\alpha_1} [1 + \exp(\underline{x}_j'\underline{\beta})]^{-n_{j0}-n_{j1}-\alpha_0-\alpha_1} \quad (4.3.1)$$

The mle $\hat{\underline{\beta}}$ is obtained via an iterative algorithm. Use of this solution method is based on the similarity between this likelihood function and one for a weighted regression model. Consider the logit $\ell_j = \ln[\pi_j/(1-\pi_j)] = \underline{x}_j'\underline{\beta}$. Under each of the three commonly assumed distributions contingency table observations, the variance of ℓ_j is approximately $\frac{n_{j0}+n_{j1}+\alpha_0+\alpha_1}{(n_{j0}+\alpha_0)(n_{j1}+\alpha_1)}$.

If we consider the linear model

$$\ell_j = \underline{x}_j'\underline{\beta} + \varepsilon_j \quad (4.3.2)$$

with the ε_j independent $\left(0, \frac{n_{j0}+n_{j1}+\alpha_0+\alpha_1}{(n_{j0}+\alpha_0)(n_{j1}+\alpha_1)}\right)$, then the iteratively reweighted least squares solution $\underline{\beta}^*$ converges to the mle $\hat{\underline{\beta}}$.

Using the Newton-Raphson algorithm to solve for $\underline{\beta}^*$, at each iteration the estimate can be written as a function of the estimate from the previous iteration. Some additional notation is needed to write the appropriate equations used in this iterative procedure. First, define

$X = (\underline{x}_1,\underline{x}_2,\ldots\underline{x}_c)'$, as the "design" matrix, in which row j contains the set of dummy variables that define cell j,

$\underline{w}_1 = (w_{11}, w_{21},\ldots w_{c1})$, where $w_{j1} = n_{j1}+\alpha_1$ is the number of observations in cell j with Y=1, and

$\underline{w} = (w_1,w_2,\ldots w_c)$, where $w_j = w_{j1}+n_{j0}+\alpha_0$ is the total number of observations in cell j.

For iteration t we have the estimates

$\hat{P}_{j1}(\underline{\beta}(t)) = \exp(\underline{x}_j'\underline{\beta}(t)) \,/\, [1 + \exp(\underline{x}_j'\underline{\beta}(t))]$ for the probability that an observation in cell j has Y=1,

$w_j\hat{P}_{j1}(\underline{\beta}(t))$ for the corresponding estimated number of observations,

$\hat{\underline{w}}_1(\underline{\beta}(t)) = (w_1\hat{P}_{11}(\underline{\beta}(t)), w_2\hat{P}_{21}(\underline{\beta}(t)), \ldots, w_c\hat{P}_{c1}(\underline{\beta}(t)))$, and

$\hat{\sigma}_j^2(\underline{\beta}(t)) = w_j\hat{P}_{j1}(\underline{\beta}(t))[1-\hat{P}_{j1}(\underline{\beta}(t))]$ for the variance of the w_1 observations in cell j, and

$\hat{V}_t$ is a diagonal matrix with j^{th} diagonal element $\hat{\sigma}_j^2(\underline{\beta}(t))$.

Then the iterative estimation equation is

$$\underline{\beta}(t+1) = \underline{\beta}(t) + [X' V_t X]^{-1} X'[\underline{w}_1 - \hat{\underline{w}}_1(\underline{\beta}(t))], \qquad (4.3.3)$$

with the estimated variance matrix of $\underline{\beta}(t+1)$ being

$$\hat{V}(\underline{\beta}(t+1)) = (X' V_{t+1} X)^{-1} . \qquad (4.3.4)$$

5. Generating Parameter Sets for Use in Imputation

For each model five sets of parameters are required, one set for use with each imputation. We would like to draw five parameter sets from the true posterior of $\hat{\underline{\beta}}$ which is proportional to

$$\prod_{j=1}^{C} [\exp(\underline{x}_j'\hat{\underline{\beta}})]^{n_{j1}+\frac{sp}{C}} [1 + \exp(\underline{x}_j'\hat{\underline{\beta}})]^{-n_{j0}-n_{j1}-\frac{p}{C}} \qquad (5.1)$$

However, this is difficult to do. Another approach is to use the large sample approximate posterior distribution

$$(\underline{\beta}-\hat{\underline{\beta}}) \sim N(\underline{0}, \hat{V}(\hat{\underline{\beta}})) . \qquad (5.2)$$

This is done by representing $\underline{\beta}$ as

$$\underline{\beta} = \hat{\underline{\beta}} + \hat{V}^{1/2}(\hat{\underline{\beta}})\ N_p(0,1) \tag{5.3}$$

where $\hat{V}^{1/2}(\hat{\underline{\beta}})$ is the Cholesky square root of $\hat{V}(\hat{\underline{\beta}})$ and $N_p(0,1)$ is a p-vector of independent N(0,1) deviates. Now values of $\underline{\beta}$ are obtained by generating sets of p independent random N(0,1) variables, placing them in a p-vector, and using the transformation (5.3). The IMSL procedure GGNSM was used to generate the normal random deviates. It uses as input the matrix $\hat{V}(\hat{\underline{\beta}})$ and produces the $\hat{V}^{1/2}(\hat{\underline{\beta}})\ N_p(0,1)$ vectors as output.

5.1 Sampling/Importance Resampling

Because of the small sample sizes used to estimate most models, the normal approximation is not necessarily close to the actual posterior distribution. For this reason it is not practical to generate 5 sets of parameters from the normal posterior and use them. Something must be done to "ensure" that the true posterior is more closely approximated. One way of doing this is via the sampling/importance resampling (SIR) procedure described in Rubin (1987).

The SIR as applied to this situation involves three steps.

1. Generate a large number of $\underline{\beta}_k$'s from the approximate normal posterior $(\hat{\underline{\beta}},\ \hat{V}(\hat{\underline{\beta}}))$.

2. Calculate the ratio of each $\underline{\beta}_k$'s true posterior to its normal posterior

$$r_k = \frac{\prod_{j=1}^{C} [\exp(\underline{x}_j'\underline{\beta}_k)]^{n_{j1}+\frac{sp}{C}} [1 + \exp(\underline{x}_j'\underline{\beta}_k)]^{-n_{j0}-n_{j1}-\frac{p}{C}}}{|\hat{V}\ (\hat{\underline{\beta}})|^{-1/2} \exp^{-\frac{1}{2}} (\underline{\beta}_k-\hat{\underline{\beta}})'\hat{V}^{-1}(\hat{\underline{\beta}})\ (\underline{\beta}_k-\hat{\underline{\beta}})} \tag{5.1.1}$$

3. Select five $\underline{\beta}_k$'s by sampling with replacement, where the probability of selecting $\underline{\beta}_k$ at each draw is proportional to r_k.

The resultant probability distribution of the $\underline{\beta}_k$'s has approximately the correct posterior distribution. The number of $\underline{\beta}_k$'s generated for a given model increased as the ratio $n._1/n._0$ diverged from one. This is because deviant values of this ratio suggest that the π_j's as a group are close to 0 or 1, and the normal approximation is less accurate at these extremes.

6. Producing Imputed Public Use Files

The actual imputation of industries and occupations was quite simple. However, the file manipulation required was not, because several computers and two locations were involved. These two phases will be discussed separately.

6.1 Imputation procedure

Consider a file of records for persons from a specified 1970 industry requiring imputation using generated parameter sets. The imputation procedure is identical for each record and each of the five imputes.

1. Compute the vector $\underline{x}$ that identifies the cell to which this record belongs.

2a. Calculate $\pi_{1k} = \exp(\underline{x}'\underline{\beta}_{1k})/[1+\exp(\underline{x}'\underline{\beta}_{1k})]$ where the subscripts denote the k^{th} imputation and the first model.

2b. Generate a uniform (0,1) random variable U_1.

2c. If $U_1 \leq \pi_{1k}$, impute the first industry. Go to the next impute.

3a. Otherwise, calculate $\pi_{2k} = \exp(\underline{x}'\underline{\beta}_{2k})/[1 + \exp(\underline{x}'\underline{\beta}_{2k})]$

3b. Generate a uniform (0,1) random variable U_2.

3c. If $U_2 \leq \pi_{2k}$, impute the second industry. Go to the next impute.

4a. Otherwise, continue this procedure with successive models until an industry is imputed. If no industry has been imputed and for the final model (f), $U_f > \pi_f$, impute the final industry (f + 1) .

For the 100% industries the procedure is even simpler. All the records for a given 1970 industry get the same 1980 industry for all imputes. This procedure is followed for all 1970 I&O codes requiring imputation.

6.2 File Manipulation

County group public use files were selected for imputation because they include the geographic identification of counties or groups of counties having populations of 250,000 or more. These public use files have two kinds of records - household and person. For imputation purposes we are interested in persons 16 years of age or older who reported having labor force activity in 1960 or later. The final product is two public use files that are identical to the original files except that five imputed industries and occupations have been added to the end of each of the appropriate records.

A file containing information for each of the records requiring imputation was created from the two public use files. It includes the variables used in the logit models and identifying information needed to match back to the original file. A large number of separate files were now created: one for each industry and occupation code requiring imputation (including major group codes), one for each aggregate occupation, one for all the 100% industries and one for all the 100% occupations. A set of such files from two 1970 county group files was created from public use tapes at UCLA and sent to the Census Bureau on tape. The total number of records needing imputed values was approximately 1.7 million.

Imputation was performed according to the procedure described in section 6.1. Tapes with the imputed records were written and matched with information from the original files to create final files of the desired form. Two copies of these files were then written onto tape, one copy of each file being sent to UCLA and one copy being sent here to the Census Bureau tape library. These files are being made available to the public through DUSD. The original tapes reside in the Integrated Statistical Library tape library.

7. Some Programming Details

Computer software was written to perform each of the three major statistical tasks: parameter estimation, parameter generation via sampling/importance resampling and imputation. All programs were written in UNISYS standard Fortran with particular subroutines from the International Mathematical and Statistical Language (IMSL) package being used where appropriate. All calculations for the industries and the 100% occupations were performed on the Census Bureau's UNISYS 1184 A-machine. The software was transferred via tape to the Census/ASA/NSF Microvax II. Updating of this software for use with occupation variables and DEC standard Fortran was then carried out on a Microvax II purchased through a grant from NSF to Dr. Rubin. Estimation, parameter generation and imputation onto the double-coded file were then performed using the NSF Microvax II and the Census/ASA/NSF Microvax II. Occupation imputation was done on a Microvax II at the UCLA Division of Biostatistics. Compilation of the imputed public use files was completed using the Statistical Standards and Methodology's Integrated Statistical Laboratory IBM 4361.

7.1 Estimation Program

The estimation program carries out several functions in addition to the iterative Newton-Raphson solution algorithm. It is written to allow for a variable number of predictor variables as well as a variable number of levels per variable. However, since the predictor variables and their levels were defined identically for all industry models, the vectors $\underline{x}_j$ of dummy variables for all cells were calculated once and stored in a file. This file was used in each model estimation. A subroutine was written for use with occupation estimation that calculates the values of $\underline{x}_j$ according to the definition of variables and levels being used for each model. The prior data amounts α_1 = sp/C and α_0 = (1-s)p/C were calculated from the distribution of original 1970 and 1980 industry code combinations.

Once this preliminary work is completed, the individual records with this 1970 code are read and counted in the appropriate cell if they are being used for the current model. When all the n_{j0} and n_{j1} have been calculated the parameters are estimated. Covergence to a solution is achieved after g iterations when for each element $\beta_i(g+1)$ of $\underline{\beta}(g+1)$,

$[\beta_i(g+1) - \beta_i(g)]/\ \beta_i(g+1) \leq .0001$ or $|\ \beta_i(g+1)\ | < .0001$.

The variance matrix of $\hat{\underline{\beta}}$ is estimated by using equation (4.3.4) and the solution vector $\underline{\beta}\ (g+1)$.

The output file from this program is used as input to the parameter generation program. It includes a list of the possible 1980 codes, $\hat{\underline{\beta}}$ and $\hat{V}(\hat{\beta})$ for each model.

7.2 Parameter Generation and Selection Program

This program has three main parts: generate a collection of parameter sets for each model, calculate the ratio of likelihoods r_k for each set and select five sets for each model. First it determines, based on n_{j0}/n_{j1}, how many sets of vectors to generate. These vectors are generated according to the procedure described in section 5. The r_k's are calculated in three steps. The approximate normal likelihood is calculated directly using $\underline{\beta}_k$, $\hat{\underline{\beta}}$ and $\hat{V}(\hat{\underline{\beta}})$. Then each observation is read and its contribution to the exact likelihood for each model is determined. Finally, the contribution of all prior data is determined.

Selection of parameters is done for one model at a time. The r_k's for a model are sorted in ascending order. A uniform $(0, \Sigma r_k)$ random variable U is generated. If j is the smallest integer such that $U \leq \sum_{k=1}^{j} r_k$, then the corresponding $\underline{\beta}_j$ is selected for use in imputation. Four additional selections are carried out by independently generating other uniform $(0, \Sigma r_k)$ random variables. The sampling at each stage is done with replacement. For each model the five selected $\underline{\beta}_j$'s are output to a file to be used in the imputation phase.

7.3 Imputation Program

This is the simplest of the three programs, since most of the calculations it uses appeared in one or both of the previous programs. A record is read and the vector $\underline{x}$ of predictor dummy variables is determined. Using $\underline{x}$, the imputation procedure described in section 6.1 is carried out for each of the five parameter sets. The input record with five imputed codes added is written to an output file. This procedure is carried out for all input records.

REFERENCES

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1984), "Simple Bayesian Methods for Logistic Regression, with Application to a Study of Inter-Class Comparability of Industry and Occupation Classification Systems," presented at the Annual Meeting of the American Statistical Association.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1988), "Simple Bayesian Methods of Logistic Regression, with Application to a Study of Inter-Class Comparability of Industry and Occupation Classification Systems," submitted for publication.

Haberman, S.J. (1977), The Analysis of Quantitative Data. Volume I: Introductory Topics. New York: Academic Press.

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, 209-245.

Rubin, D.B. (1978), "Multiple Imputations in Sample Surveys," Proceedings of the Survey Research Methods Section of the American Statistical Association, 20-34.

Rubin, D.B. (1980), Handling Nonresponse in Sample Surveys by Multiple Imputation, Bureau of the Census Monograph, Washington, D.C.: U.S. Government Printing Office.

Rubin, D.B. (1983), "Progress Report on Project for Multiple Imputation of 1980 Codes," University of Chicago.

Rubin, D.B. (1987), "A Noninterative Sampling/Importance Resampling Alternative of the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm, " (Comment on "The Calculation of Posterior Distributions by Data Augmentation"). Journal of the American Statistical Association, 82, 543-546.

Rubin, D.B. (1988), Multiple Imputation for Nonresponse in Surveys, New York: Wiley.

Rubin, D.B. and Schenker, N. (1987), "Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior," Sociological Methodology, 17, 131-144.

Schenker, N., Treiman, D.J., and Weidman, L. (1988), "Multiple Imputation of Industry and Occupation Codes for Public-Use Data Files," Proceedings of the Survey Research Methods Section of the American Statistical Association (to appear).

Schenker, N., Treiman, D.J., and Weidman, L. (1989), "Analyses of Public-Use Data with Multiply-Imputed Industry and Occupation Codes," submitted for publication.

Subcommittee of Comparability of Occupation Measurement (1983), "Alternative Methods for Effecting the Comparability of Occupation Measurement over Time," report to the Social Science Research Council Advisory and Planning Committee on Social Indicators and the U.S. Bureau of the Census, Washington, D.C.: Social Science Research Council.

Treiman, D.J., Bielby, W.T., and Cheng, M.T. (1988) "Evaluating a Multiple-Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard," Sociological Methodology, 18, 309-345.

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.

U.S. Bureau of the Census (1971), _1970 Census of Population Classified Index of Industries and Occupations_, Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1980), _1980 Census of Population Classified Index of Industries and Occupations_, Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1972), _Public Use Samples of Basic Records from the 1970 Census: Description and Technical Documentation_, Washington, D.C.: U.S. Government Printing Office.

Figure 1

Industry and Occupation Imputation Project Overview

Parameter Estimation and Imputation Evaluation

Data File:

127,125 case double-coded file with 1970 & 1980 I & O codes	Sort into Industry and Occupation Subfiles	

Tasks:

Model estimation for industry	Add prior data to each cell and estimate parameters via Newton-Raphson	
Parameter generation and selection for industries	Generate parameter sets from asymptotic normal posteriors	Calculate ratios of likelihoods r_k and select 5 parameter sets
Imputation of industries on double-coded files	Using selected parameter sets impute 5 1980 Industry codes onto each record	
Evaluation of industry imputation	Compare various distributions of hand-coded and imputed 1980 codes	

Repeat Tasks for Occupation

Multiple Imputation of Public Use Files

Data File:

1970 Public Use File (PUMS)	Select person records with proper ESR and age, retaining needed information	Sort into Industry and Occupation subfiles

Tasks:

PUMS imputation	Using selected parameter sets, impute 5 1980 I & O codes onto each record
Final file creation	Match imputed records back to original PMUS and create a final PUMS that is just like the original, with I & O imputes added

Table 1

1970 Industry Codes: Treatment for Imputation

Industries with Logit Models

017	178	279	397	558	669	778	889
018	179	287	398	567	677	779	897
019	187	288	407	568	678	787	907
027	188	297	408	569	679	797	917
028	189	298	409	587	687	798	927
047	197	307	417	588	688	807	937
049	198	317	418	607	697	808	
057	199	318	427	608	698	809	
067	207	319	429	609	707	828	
068	208	327	447	617	708	829	
069	209	328	448	618	709	837	
108	219	329	467	619	717	838	
109	227	338	468	627	718	839	
118	229	339	477	628	728	847	
119	237	347	478	629	729	848	
127	238	348	479	637	737	849	
138	239	349	507	638	738	857	
139	247	357	508	639	747	858	
147	248	358	509	647	748	859	
149	257	368	527	648	749	867	
157	258	369	528	649	757	877	
158	259	377	529	657	758	878	
167	268	379	537	658	759	887	
168	269	387	539	667	769	888	
169	278	389	557	668	777		

100% Industries with Corresponding 1980 Codes

1970	1980
048	041
077	060
107	230
128	252
137	261
148	272
159	290
177	310
228	360
249	381
289	120
299	130
308	140
309	141
337	162
359	190
367	191
378	201
388	220
419	420
428	422
449	442
469	461
538	522
559	531
689	681
727	721
739	740
788	780
789	782
868	860
869	872
879	870
879	870
997	992

Major Industry Groups

Major Group Codes	Industry Code Range
029	017-028
058	047-057
078	067-077
267	107-259
399	268-397
499	407-479
599	507-588
699	607-698
719	707-718
767	727-759
799	769-798
817	807-809
899	828-897
947	907-937

Table 2

1980 Occupation Codes: Treatment for Imputation

Occupations with Logit Models

001	174	314	433	612	754
002	180	315	436	613	755
003	181	320	441	614	760
006	183	321	442	621	761
011	184	323	445	623	762
012	185	325	450	625	764
013	190	326	452	630	770
014	191	330	454	633	780
023	194	331	461	636	785
025	195	332	470	640	801
031	201	333	471	642	802
032	202	341	472	643	821
045	205	342	473	644	822
053	210	343	475	645	823
055	212	345	480	650	901
056	215	355	482	651	902
064	216	360	485	652	903
065	220	361	486	653	910
074	222	364	492	656	911
076	223	372	495	660	912
080	225	374	510	661	914
083	226	375	511	662	915
085	230	376	512	663	916
086	231	381	514	664	921
090	233	382	520	666	922
091	235	384	522	672	926
101	245	385	530	674	932
112	262	390	533	680	933
122	264	391	534	681	934
124	265	394	535	690	935
126	271	395	542	692	950
135	281	402	545	694	960
140	282	404	550	695	961
141	283	405	551	703	962
142	284	410	552	705	964
143	285	412	554	706	980
144	301	415	560	711	981
145	303	421	575	713	982
152	305	422	602	714	984
153	310	423	604	715	
164	312	424	610	740	
165	313	430	611	751	

Equal Probability Occupations		Major Occupation Groups	
		Major Group Codes	**Occupation Code Range**
054		196	001-195
747		246	201-245
521		296	260-285
945		396	301-395
		586	401-580
		696	601-695
		726	701-785
		796	740-785
		806	801-802
		846	821-824
		976	901-965
		986	980-984

Aggregate Occupations

1004 aggregate
004
005

1042 aggregate
042
044

1063 aggregate
063
506

1075 aggregate
075
923

1100 aggregate
100
311

1104 aggregate
104
113

1150 aggregate
150
151
154
155
162
173

1161 aggregate
161
605

1163 aggregate
163
170

1221 aggregate
221
701

1270 aggregate
270
363

1401 aggregate
401
563

1403 aggregate
403
481
491
540
571

1420 aggregate
420
440

1434 aggregate
434
515

1446 aggregate
446
626

1453 aggregate
453
516

1503 aggregate
503
546

1561 aggregate
561
562

1620 aggregate
620
641

1704 aggregate
704
931

1753 aggregate
753
763

1924 aggregate
924
925

1942 aggregate
942
952

100% Occupations with Corresponding 1980 Codes

1970	1980
010	048
015	045
020	046
021	047
022	258
024	159
026	159
030	179
033	165
034	066
035	068
036	067
043	074
051	075
052	075
061	089
062	085
071	088
072	086
081	204
082	205
084	446
093	167
094	168
095	173
096	169
102	136
103	113
105	115
110	116
111	127
114	118
115	135
116	119
120	123
121	125
123	137

1970	1980
125	139
130	144
131	149
132	145
133	147
134	148
156	159
171	228
172	233
175	187
182	193
192	197
193	198
203	028
211	018
213	035
224	017
240	014
260	256
261	284
266	278
334	366
344	345
350	309
362	384
370	313
371	313
383	357
392	368
411	564
413	657
416	569
425	185
426	678
431	576
435	649
443	658
444	674
455	824
456	824
462	639

1970	1980
483	518
484	538
501	768
502	544
504	675
505	773
523	587
525	695
531	734
536	654
543	789
572	217
580	783
601	593
603	615
615	573
622	766
624	799
631	686
634	888
635	723
665	784
670	749
671	739
673	739
710	824
712	825
750	869
752	498
824	479
913	444
940	469
941	469
943	454
944	458
953	464
954	467
963	423
965	423
983	403
991	909

Table 3

Independent Variables and Number of Parameters

Variable Name	# Parameters
Used with both industry and occupation	
Sex	1
Race	1
Sex x Race	1
Age	3
Sex x Age	1
Class of Worker [1]	2,1
Hours Worked	1
Weeks Worked	1
Education	1, variable
Used with industry only	
Metro/Non-metro	1
Region	2
Used with Occupation only	
Earnings	3
Earnings x Sex	3
Earnings x Race	3
Earnings x Sex x Race	3
1970 industry [2]	variable
1970 occupation [3]	variable

Note:
1. If there are two numbers of # parameters, the first is for industry and the second for occupation.
2. 1970 industry is only used for specified occupations.
3. 1970 occupation is used where aggregates of occupations are modeled together.

Table 4

Definitions of Independent Variables and Categories

Independent Variables	Double-coded	PUMS[1]
Sex		
Male	Sex=M	P6=0
Female	Sex=F	P6=1
Race		
Black	Race=N	P7=1
Other	Race=W,0	P7=0,2-8
Age		
AGE 0	Age=16-24	Same, using P11 for Age
AGE 1	Age=25-39	
AGE 2	Age=40-59	
AGE 3	Age=60+	
Class of Worker		
Industry		
COW 0	COW=0	Same, using P30 for COW
COW 1	COW=1-3	
COW 2	COW=4-6	
Occupation		
COW 0	COW=0-3	Same, using P30 for COW
COW 1	COW=4-6	
Metro		
Industry Only		
Metro	States 10, 44 Other states SMSA=1-9998	County group (H7-11) in list on pp. 123-126
Nonmetro	SMSA=0	Complement of Metro

Education

Industry		
Not College	HGC=E,H	Complement of College
College	HGC=C	HGA=16, HGA=15 and HGC=1

Occupation

Variable Number of Levels with years of education defined by		
	1 if HGC=K	0 if HGA=0,1,2
	HGA if HGC=E	HGA-2 if HGC=1 and $3 \leq$ HGA ≤ 20
	HGA+8 if HGC=H	HGA-3 if HGC=0,2,3 and $3 \leq$ HGA ≤ 20
	HGA+12 if HGA=C	

Hours Worked

Part time	HW=0,1,2	HW=0,1,2
	HW=8 and ESR=2	HW=8 and ESR=3
Full time	HW=3-7	HW=3-7
	HW=8 and ESR=1	HW=8 and ESR=1,2

Weeks Worked

Part year	WW=0,1,2	WW=0,1,2
	WW=6 and ESR=2	WW=6 and ESR=3
Full year	WW=3,4,5	WW=3,4,5
	WW=6 and ESR=1	WW=6 and ESR=1,2

Region

Industry Only		Defined by H7-11=county group values
East	States in New England, Mid Atlantic, East North Central Divisions	00101-01600,05101-06100, 06201-06206,06301-07401,07403-07500,07601-08000,08002-08200, 10101-10102,10107-10200
South	States in South Atlantic, East and West South Central Divisions	01601-05100,06101-06200,06207-06300,10202-10300,10401-12801

West	States in West North Central Mountain, Pacific Divisions	07402,07501-07600,08001,08201-10100,10103-10106,10201,10301-10400,12802-14902

Earnings
Occupation only
by quartiles

	Largest values defined = $99,000	Same, using P37-39 Largest value defined = $50,000 (largest upper quartile < $50,000)

Note: 1. All references to variable names and pages refer to Public Use Samples of Basic Records From the 1970 Census: Description and Technical Documentation, U.S. Bureau of the Census, Washington, D.C., 1972.