BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-87/09

REPORT ON MISSING DATA IN THE 1986 TEST OF ADJUSTMENT RELATED OPERATIONS

by

Nathaniel Schenker
Undercount Research Staff
Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended by: Kirk M. Wolter

Report completed: April 8, 1987

Report issued: April 8,1987

This report discusses missing data problems in the estimation of undercount for the 1986 Test of Adjustment Related Operations (TARO). It is assumed that the reader is familiar with the design of TARO and the basic ideas of dual-system estimation, as described in Diffendal (1987) and Wolter (1986).

Section 1 describes the data needed for undercount estimation in TARO. Sections 2-5 discuss the types of missing data that occurred in TARO, the extent to which they occurred, and the methods used to handle them. Finally, Section 6 presents undercount estimates under several alternative treatments of missing data.

1. Data Used for Undercount Estimation

The dual-system estimator of the population size that was used in TARO is written

$$DSE = N_p(CEN-SUB-EE)/M,$$

where N_p is the weighted number of people in the P sample, CEN is the unadjusted census count, SUB is the number of whole-person substitutions in the census, EE is a weighted estimate of the number of erroneous enumerations and unmatchable persons in the census, and M is the weighted number of matches between the P sample and census. Census data provide CEN and SUB, whereas P- and E-sample data provide N_p, EE, and M.

In TARO, dual-system estimates were computed within post-strata based on person and household characteristics. (See Diffendal 1987 for a description of the post-strata.) Thus the P- and E-sample data needed for undercount estimation were the match/nonmatch (M/N) status for each P-sample person, the correct enumeration/erroneous enumeration (CE/EE) status for each E-sample person, and household and person characteristics for each person in both samples.

After the PES was conducted, a two-way match was performed between the PES and census. The result was that most P-sample cases were assigned an M/N status, all E-sample matches were classified as CE's, some of the E-sample nonmatches were classified as EE's (e.g., duplicates and unmatchables), and the remaining E-sample nonmatches were followed up to determine their CE/EE statuses. After followup, the unresolved P- and E-sample cases were missing their M/N and CE/EE statuses, respectively.

2. Household Noninterviews

Of the 5,935 nonvacant households in the TARO P sample, 32 (0.54%) were classified as noninterview households. These included 11 last-resort households, that is, households with incomplete interviews with a respondent from outside the household. The occurrence of household noninterviews resulted in missing data on the number of people in each household, person and household characteristics, and M/N statuses.

Household noninterviews were handled by adjusting the sampling weight for every person in an interviewed household as follows. Within each block, the sampling weight for every person was the same. This sampling weight was inflated by the inverse of the completed-interview rate for the block. The noninterview weighting adjustment basically assumes that the distributions of people, characteristics, and M/N statuses for households not interviewed within a block are the same as for households interviewed.

It is possible that the data obtained for households by proxy interviews (that is, completed interviews with neighbors or landlords) are of sufficiently low quality that such households should be classified as noninterview households. The quality of data from the 189 proxy interviews in TARO is discussed in Section 3, and some undercount estimates with proxy interviews treated as noninterviews are presented in Section 6.

3. Missing Characteristics in the P and E Samples

The variables used for post-stratification in TARO

(Diffendal 1987) included the housing unit variable TENURE (1 = owned, 2 = rented or occupied without payment) and the person variables SEX (1 = male, 2 = female), AGE (1 = 0-14, 2 = 15-29, 3 = 30-44, 4 = 45-64, 5 = 65+), and RACE (1 = Hispanic, 2 = Asian non-Hispanic, 3 = Other). In addition, the housing unit variable STRUCTURE (1 = single-unit, 2 = multiunit) was used in handling missing P-sample M/N statuses and missing E-sample CE/EE statuses (see Sections 4 and 5).

Table 1 displays the missing characteristic data counts for the entire P and E samples and for P-sample proxy interviews. For the P and E samples, the highest missing data rate was 7.0% for E-sample RACE, with all other rates being 3.5% or lower. The missing data rates for P-sample proxy interviews were all several times higher than those for the entire P sample, although only TENURE (20.2%) had a rate higher than 10%.

Missing characteristics for each of the samples (P and E) were imputed by a hot-deck method involving two passes through the geographically sorted data. On the first pass, TENURE, STRUCTURE, and RACE were imputed using the most recent observed data, because of the presumed strong relation between these variables and geography. On the second pass, SEX and AGE were imputed at random from distributions tabulated during the first pass using all observed data.

For the first-pass sequential imputations, persons were grouped into households. Whenever TENURE was observed but STRUCTURE was missing, the most recent household having both variables observed and the same value of TENURE as the household in question was used to provide a value of STRUCTURE to impute. The situation with STRUCTURE observed and TENURE missing was treated analogously. When both TENURE and STRUCTURE were missing, the values for the most recent household with both variables observed were substituted. Whenever RACE was missing for any person in a household, the most recent household with any observed values of RACE (which may have been the household in question) was used to compute a RACE distribution; the missing values were then imputed randomly from this distribution.

The imputation of SEX and AGE during the second pass through the data controlled for several factors. Thus observed SEX and AGE distributions were tabulated during the first pass for several different categories. Specifically, the imputation of SEX controlled for whether the person in question lived in a single-person or multiperson household; for multiperson households, the imputation also controlled for the relationship of the person in question to the head of household. The imputation of AGE controlled for whether the household was single-person or multiperson as well as marital status and (for multiperson households) relationship to the head of household and age of the head of household.

Tables 2 and 3 compare the observed and imputed distributions of characteristics in the P and E samples, respectively.

4. Missing M/N Statuses in the P Sample

Of the 19,552 P-sample cases resulting from completed interviews, 161 (0.8%) were missing M/N statuses for dual-system estimation. All but three of the unresolved cases fell into two broad categories: 105 cases for which matching was not attempted due to incomplete names and/or insufficient characteristics; and 53 movers for whom there were problems obtaining or geocoding a Census Day address or finding the census questionnaire for the Census Day address.

After all missing characteristics were imputed using the methods described in Section 3, a match probability was imputed for each unknown M/N status. The contribution of the unresolved cases to the denominator of the dual-system estimate (Section 1) was the weighted sum of the imputed probabilities. Because imputed probabilities represent a degree of uncertainty about the missing M/N statuses, the probabilities can be used to obtain a variance due to imputation. Current research to be presented at the 1987 meeting of the American Statistical Association is developing methods of calculating this imputation variance.

The following logistic regression approach was used to impute match probabilities. Let X denote a vector of predictors, Y = M or N, and p = Pr(Y=M|X). The parameter vector β of the logistic regression model

$$\text{logit}(p) = \log[p/(1-p)] = X'\beta$$

was estimated from the data for the resolved cases using the Bayesian techniques described in Clogg, Rubin, Schenker, Schultz, and Weidman (1986) and Rubin and Schenker (1987). Then for unresolved case j, with $X=x_j$, the imputed match probability was

$$\hat{p}_j = \text{logit}^{-1}(x_j'\hat{\beta}) = \exp(x_j'\hat{\beta})/[1 + \exp(x_j'\hat{\beta})] \text{ ,}$$

where $\hat{\beta}$ denotes the estimate of β. The background variables used to define X for TARO were TENURE, STRUCTURE, SEX, AGE, and RACE, as well as variables indicating regular interview versus proxy interview and mover versus nonmover.

Of the 19,391 resolved P-sample cases, 17,018 (87.8%) were matches. The (unweighted) sum of the 161 imputed match probabilities was 124.66; thus the imputed match rate was 77.4%. At a February 1987 workshop on the undercount at Harvard University, it was suggested that indicator variables for the six sampling strata (Diffendal 1987) be included in X. The result of this refinement is a sum of imputed match probabilities equal to 124.50 (77.3%). The very minor effect of this change on estimates of the undercount is demonstrated in Section 6.

5. Missing CE/EE Statuses in the E Sample

Of the 20,976 cases in the E sample, 3,714 were sent or should have been sent to followup. After followup, 979 cases (4.7% of total, 26.4% of followup) had missing CE/EE statuses. All but nine of the unresolved cases fell into four broad categories: 498 cases that should have been sent to followup but were not; 257 cases in which the respondent to the followup interview did not know the person in question; 137 cases for which the interview yielded insufficient information to determine a CE/EE status; and 78 cases for which there were followup noninterviews.

Missing CE/EE statuses in the E sample were handled by imputing a probability of erroneous enumeration for each unresolved case. The contribution of the unresolved cases to the EE term in the numerator of the dual-system estimate (Section 1) was the weighted sum of the imputed probabilities. The imputation procedure is analogous to that used for P-sample M/N statuses with one major change: Since missing CE/EE statuses resulted solely from followup, only the resolved cases from followup were used in estimating the logistic regression. The background variables used to define X for the logistic regression were TENURE, STRUCTURE, SEX, AGE, and RACE, along with variables indicating whether the census questionnaire for the person's household was returned by mail and whether the entire household or only part of the household was not matched before followup.

Of the 17,262 non-followup cases, 278 (1.61%) had status EE. There were 2,735 resolved followup cases, of which 82 (3.0%) had status EE. The (unweighted) sum of the 979 imputed probabilities was 21.93 (2.2%). When indicator variables for the sampling strata are included in X, the sum changes to 23.58 (2.4%). As with the P sample, this change has a very minor effect on estimates of the undercount; see Section 6.

6. Undercount Estimates Under Alternative Treatments of Missing Data

This section examines the effects of alternative treatments of missing data on estimated undercount rates for the three categories of race defined by the variable RACE (Hispanic, Asian non-Hispanic, and Other). For a given treatment and race category, let $\hat{N}$ be the sum of the dual-system estimates over all post-strata corresponding to the race category and let N_c be the sum of the unadjusted census counts over the post-strata. The estimated undercount rate is then $100(1 - N_c/\hat{N})\%$.

Consider first the suggestion discussed in Sections 4 and 5 to include indicators of the sampling strata as predictors in the P- and E-sample logistic regressions for imputing match and erroneous enumeration probabilities. The TARO estimated undercount rates, which were obtained without using these predictors, are 9.85% for Hispanics, 7.32% for Asian non-Hispanics, and 6.24% for Others. When indicators of the sampling strata are used, the estimates change to 9.82% for Hispanics, 7.31% for Asian non-Hispanics, and 6.21% for Others. The largest difference due to including the sampling stratum indicators is only 0.03%. For all of the alternative treatments to be considered, however, this refinement is used because it is in principle more correct.

6.1 Treatments that Lower the Estimated Undercount

The match rate for the 375 resolved P-sample proxy cases was 78.9% as opposed to the overall P-sample rate of 87.8%. While it may be true that proxy cases were actually captured in the census less frequently than others, it is possible that part of the difference in the match rates is due to missing and/or incorrect proxy data (see Section 3). A conservative treatment would be to classify the 189 proxy interviews as household noninterviews and apply the weighting adjustment described in Section 2; this would essentially assign proxy cases the same match rate as nonproxy cases. Note that when all proxy interviews are classified as noninterviews, an indicator of proxy/nonproxy status is no longer included in the logistic regression model for imputing match probabilities.

The match rate for the 277 resolved P-sample movers was 66.1%. It is generally believed that movers are captured in the census at a lower rate than nonmovers, but it may be that the low match rate for movers is partly due to difficulites inherent in matching movers, such as problems in obtaining a correct Census Day address. A conservative treatment would be to classify all cases for movers as unresolved and then impute match probabilities for unresolved cases using a logistic regression model that does not include mover/nonmover status as a predictor. This would essentially assign movers the same match rate as nonmovers.

Of the 979 unresolved E-sample cases, 257 had the followup interview code W1, meaning that the respondent did not know the person in question. Since a code of W1 could indicate that the person in question was fictitious, all W1's were reviewed by experienced matching personnel. Any case that showed evidence (such as a note from the interviewer) of possibly being fictitious was marked; there were 118 such cases. An alternative treatment to that used in TARO would be to assign a status of EE to all of these cases. This would raise both the observed and imputed EE rates.

Table 4 displays the undercount estimates by race category for the 2x2x2 factorial design with the factors being whether or not alternative treatments are used for proxy interviews, movers, and W1's. The ranges between the lowest and highest estimated undercount rates are 1.31% for Hispanics, 1.41% for Asian non-Hispanics, and 0.43% for Others.

Note that for each race category, there is not much interaction between the treatments of proxy interviews, movers, and W1's. In fact, the following additive model can be used to predict the entries in Table 4 for each race category:

$$\hat{Y} = \hat{\alpha}_0 + I_p\hat{\alpha}_p + I_m\hat{\alpha}_m + I_w\hat{\alpha}_w \ , \qquad (1)$$

where $\hat{Y}$ is the predicted estimate of the undercount rate, I_p, I_m, and I_w are the treatment indicators (1=alternative, 0=TARO) for proxy interviews, movers, and W1's, respectively, and $\hat{\alpha}_0$, $\hat{\alpha}_p$, $\hat{\alpha}_m$, and $\hat{\alpha}_w$ are given in Table 5. The parameter α_0 is the estimated undercount rate when no alternative treatments are used; α_p, α_m, and α_w are the effects of using alterative treatments for proxy interviews, movers, and W1's, respectively. The largest residual when equation (1) is used to predict the entries in Table 4 is 0.02%.

6.2 A Treatment that Raises the Estimated Undercount

Because TARO was confined to one small area in the United States, data for people who moved outside the test site between Census Day and the PES could not be obtained. The omission of these outmovers from TARO undercount estimation was equivalent to assuming that they had the same capture rate in the census as the included cases. This was a conservative assumption, since movers are generally believed to have a lower capture rate than nonmovers.

An alternative procedure that might indicate the effect of including outmovers in the estimation would be to include the 409 people who moved into the test site between Census Day and the PES. The M/N statuses for these inmovers would be considered missing and would have probabilities imputed for them.

The treatments yielding the highest and lowest estimates in Table 4 have been applied to the TARO data with inmovers included; the results are displayed in Table 6. Note that the lower estimated undercount rates in Table 6 (obtained using the alternatives to the TARO treatments for proxy interviews, movers, and W1's) are all within 0.04% of the corresponding estimates in Table 4. This result is expected, since the addition of cases having an imputed match rate that is approximately the same as the overall match rate should not affect the estimates much. The higher etimates in Table 6 are larger than the corresponding estimates in Table 4 by 0.34% for Hispanics, 0.50% for Asian non-Hispanics, and 0.38% for Others.

6.3 Summary and Discussion

To summarize, the lowest and highest estimated undercount rates obtained using alternative treatments of missing data are 8.50% and 10.16% for Hispanics, 5.86% and 7.81% for Asian non-Hispanics, and 5.81% and 6.59% for Others. The TARO estimates for the three race categories are 9.85%, 7.32%, and 6.21%, respectively.

Note that the alternatives to the TARO procedures for handling proxy interviews and movers that were described in Section 6.1 are extreme in the sense that they essentially assume that proxy and mover cases have the same capture rates in the census as other cases. It is suspected that the correct treatments of proxy interviews and movers lie somewhere between the TARO treatments and the alternatives discussed here.

ACKNOWLEDGEMENTS

I would like to thank Errol Rowe of the Statistical Research Division for writing the computer programs that processed the raw TARO data before imputation, and Robert O'Brien of the Decennial Operations Division for computing the dual-system estimates used in Section 6. I am also grateful to the members of the workshop on the undercount of the Harvard University Statistics Department for helpful discussions on handling missing data in undercount estimation.

REFERENCES

Clogg, C.C., D.B. Rubin, N. Schenker, B. Schultz, and L. Weidman (1986), "Simple Bayesian Methods for Logistic Regression," American Statistical Association Meeting, August 1986, Chicago, Illinois.

Diffendal, G. (1987), "1986 Test of Adjustment Related Operations Procedures and Methodology," Joint Advisory Committee Meeting, April 1987, Rosslyn, Virginia.

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

Wolter, K.M. (1986), "Some Coverage Error Models for Census Data," Journal of the American Statistical Association, 81, 338-346.

Table 1

Missing Characteristic Data Counts (% in Parentheses) for the Entire P and E Samples and for P-Sample Proxy Interviews

Variable	P Sample (19,552 persons)	E Sample (20,976 persons)	P-Sample Proxy (430 persons)
TENURE	690 (3.5)	154 (0.7)	87 (20.2)
STRUCTURE	459 (2.3)	343 (1.6)	38 (8.8)
SEX	418 (2.1)	82 (0.4)	18 (4.2)
AGE	137 (0.7)	432 (2.1)	18 (4.2)
RACE	155 (0.8)	1463 (7.0)	17 (4.0)

Table 2

Observed and Imputed Distributions (in %) of Characteristics in the P Sample

	TENURE	
	1	2
Observed	47.5	52.5
Imputed	40.9	59.1

	STRUCTURE	
	1	2
Observed	78.6	21.4
Imputed	74.1	25.9

	SEX	
	1	2
Observed	48.8	51.2
Imputed	51.2	48.8

	AGE				
	1	2	3	4	5
Observed	28.5	27.7	19.4	16.4	8.0
Imputed	27.0	24.8	19.7	19.0	9.5

	RACE		
	1	2	3
Observed	75.5	9.3	15.2
Imputed	76.8	9.7	13.5

Table 3

Observed and Imputed Distributions (in %)
of Characteristics in the E Sample

	TENURE 1	TENURE 2
Observed	45.8	54.2
Imputed	44.8	55.2

	STRUCTURE 1	STRUCTURE 2
Observed	74.7	25.3
Imputed	70.6	29.4

	SEX 1	SEX 2
Observed	48.9	51.1
Imputed	56.1	43.9

	AGE 1	AGE 2	AGE 3	AGE 4	AGE 5
Observed	26.7	27.8	20.9	15.8	8.8
Imputed	20.1	23.6	21.3	19.2	15.7

	RACE 1	RACE 2	RACE 3
Observed	74.5	10.1	15.4
Imputed	73.1	12.1	14.8

Table 4

Estimated Undercount Rates (in %) by Race Under Alternative Treatments of P-sample Proxy Interviews, P-sample Movers, and E-sample W1's

Treatment Indicator (1=alterantive, 0=TARO)					
Proxy	Mover	W1	Hispanic	Asian non-Hispanic	Other
0	0	0	9.82	7.31	6.21
0	0	1	9.30	6.76	5.83
0	1	0	9.33	7.24	6.19
0	1	1	8.80	6.69	5.81
1	0	0	9.55	6.52	6.24
1	0	1	9.03	5.96	5.86
1	1	0	9.04	6.45	6.22
1	1	1	8.51	5.90	5.84

NOTE: Indicators of the sampling strata were used as predictors in the logistic regressions for imputing match and erroneous enumeration probabilities.

Table 5

Parameter Estimates for the Additive Model (1) for Predicting the Estimated Undercount Rates in Table 4

	Hispanic	Asian non-Hispanic	Other
$\hat{\alpha}_0$	9.82	7.31	6.21
$\hat{\alpha}_p$	-0.28	-0.7925	0.03
$\hat{\alpha}_m$	-0.505	-0.0675	-0.02
$\hat{\alpha}_w$	-0.525	-0.5525	-0.38

Table 6

Estimated Undercount Rates (in %) by Race When Inmovers Are Included in the Data

Treatment Indicator (1=alternative, 0=TARO) Proxy	Mover	W1	Hispanic	Asian non-Hispanic	Other
0	0	0	10.16	7.81	6.59
1	1	1	8.50	5.86	5.81

NOTE: Indicators of the sampling strata were used as predictors in the logistic regressions for imputing match and erroneous enumeration probabilities.