HANDLING MISSING DATA IN THE 1986 TEST OF ADJUSTMENT RELATED OPERATIONS

Nathaniel Schenker, Bureau of the Census*
Statistical Research Division, Washington, DC 20233

The 1986 Test of Adjustment Related Operations (TARO) was a test of census coverage error estimation and adjustment applied to the 1986 Census of Central Los Angeles County, California. This paper discusses the missing data problems that arose in the estimation of coverage error for TARO and the methods that were used to handle them. For a full description of TARO, see Diffendal (1987).

Section 1 gives a brief description of how coverage error was estimated 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 coverage error estimates under several alternative treatments of missing data and other problem cases.

1. Estimating Census Coverage Error

Census coverage error was estimated in TARO using data from a post-enumeration survey (PES) of people in the census site. First a sample of blocks in the site was drawn. Then each housing unit in the sample blocks was surveyed to determine its occupants on Census Day, its occupants at the time of the PES and where they lived on Census Day, and the characteristics of the occupants.

Two samples were used in estimating census coverage error. The P (population) sample was composed of the people who lived in the PES sample blocks at the time of the PES. An attempt was made to match each P-sample person to a person enumerated in the census to determine whether the P-sample person had been enumerated; the match rate was used essentially to estimate the capture rate of the census for the entire population. The E (enumeration) sample was composed of the people who were enumerated in the census as having lived in the PES sample blocks; this sample was used to estimate the number of erroneous enumerations (e.g., fictitious enumerations and duplicates) and unmatchable persons (e.g., persons for whom no names were reported) in the census. An attempt was made to match each E-sample person to a person in the PES. Each E-sample match was considered a correct enumeration since the PES indicated that the person should have been enumerated. Each E-sample nonmatch was followed up to determine whether it was an erroneous enumeration or a correct enumeration that was missed in the PES (which is not itself assumed to have perfect coverage).

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

$$DSE = N_p(CEN\text{-}SUB\text{-}EE)/M, \qquad (1)$$

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. The dual-system estimator can be thought of as inflating the estimated number of correct and matchable census enumerations (CEN-SUB-EE) by the inverse of the estimated census capture rate (M/N_p).

The theory of dual-system estimation assumes that for both the census and the PES, the probability of capture is constant across all people in the population (Wolter 1986). To make this assumption more realistic in TARO, separate dual-system estimates were computed within post-strata based on person and household characteristics.

To summarize, the data needed for coverage error estimation were the match status (match vs. nonmatch) for each P-sample person, the enumeration status (correct vs. erroneous) for each E-sample person, and person and household characteristics for each person in both samples.

2. P-Sample Household Noninterviews

Occasionally, a PES interviewer was unable to obtain an interview from an occupied housing unit; this occurred, for example, when the occupants refused to respond. Of the 5,935 nonvacant housing units in the TARO P sample, 32 (0.5%) were classified as having household noninterviews. The occurrence of household noninterviews resulted in missing data on the number of people in each household, person and household characteristics, and match statuses.

The block-sample design of the PES afforded a simple way to handle P-sample household noninterviews. Within each sample block, the sampling weights of the noninterview households were redistributed across the interviewed households. The noninterview weighting adjustment basically assumes that the distributions of people, characteristics, and match statuses for households not interviewed within a block are the same as for households interviewed.

It is possible that the data obtained for a household by proxy interview (that is, a completed interview with someone outside the household) are of sufficiently low quality that such a household should be classified as a noninterview household. The quality of data from the 189 proxy interviews in TARO is discussed in Section 3, and some coverage error estimates with proxy interviews treated as noninterviews are presented in Section 6.

3. Missing Characteristics in the P and E Samples

Even when an interview was obtained for a P-sample household, the data on person and household characteristics were sometimes incomplete. Incomplete data on characteristics also occurred in the E sample.

The variables used in post-stratification for 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 match statuses and missing E-sample enumeration statuses (see Sections 4 and 5).

Table 1 displays the missing characteristic data counts for the entire P and E samples and for cases coming from 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 cases 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 data after the data had been sorted geographically. 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 and STRUCTURE were missing for a household, the most recent household with complete data on these variables was used to provide imputed values. When only TENURE was missing, its value was imputed from the most recent household having complete data and the same value of STRUCTURE as the household in question; missing values of STRUCTURE were imputed analogously. 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.

4. Missing Match Statuses in the P Sample

Of the 19,552 P-sample cases resulting from completed interviews, 161 (0.8%) were missing match statuses for dual-system estimation. All but three of these 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 between Census Day and the PES for whom there were problems specifying 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 match status. The contribution of the unresolved cases to the M term of the dual-system estimate (1) was the weighted sum of the imputed probabilities. Because imputed probabilities represent a degree of uncertainty about the missing match statuses, the probabilities can be used to obtain a variance due to imputation; this is also true of the imputed erroneous enumeration probabilities discussed in Section 5. Current research 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 = match or nonmatch, and $\rho = Pr(Y=\text{match}|X)$. The parameter vector β of the logistic regression model

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

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

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

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 between Census Day and the PES.

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 census coverage error is demonstrated in Section 6.

5. Missing Enumeration Statuses in the E Sample

Of the 20,976 cases in the E sample, 3,714 were followed up or should have been followed up. After followup, 979 cases (4.7% of total, 26.4% of followup) had missing enumeration statuses. All but nine of these unresolved cases fell into four broad categories: 498 cases that should have been followed up 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 an enumeration status; and 78 cases for which there were followup noninterviews.

Missing enumeration 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 of the dual-system estimate (1) was the weighted sum of the imputed probabilities. The imputation procedure was analogous to that used for P-sample match statuses with one major change: Since missing enumeration 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 (1.6%) were classified as erroneous enumerations or unmatchable. There were 2,735 resolved followup cases, of which 82 (3.0%) were classified as erroneous enumerations. 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 coverage error; see Section 6.

6. Estimates of Coverage Error Under Alternative Treatments of Missing Data and Other Problem Cases

This section examines the effects of alternative treatments of missing data and other problem cases on estimates of coverage error 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 (between Census Day and the PES) 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 classify the 118 cases as resolved erroneous enumerations before imputation. This would raise both the observed and imputed rates of erroneous enumeration.

Table 2 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 2 for each race category:

$$\hat{Y} = \hat{\alpha}_0 + I_p\hat{\alpha}_p + I_m\hat{\alpha}_m + I_w\hat{\alpha}_w , \qquad (2)$$

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 3. 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 alternative treatments for proxy interviews, movers, and W1's, respectively. The largest residual when equation (2) is used to predict the entries in Table 2 is 0.02%.

6.2 A Procedure that Raises the Estimated Undercount

Because TARO was confined to one small area in the United States, data for people who moved out of the test site between Census Day and the PES, which should have been included in the coverage error estimation, could not be obtained. The omission of these outmovers from TARO 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.

There were 409 people who moved into the test site between Census Day and the PES. These inmovers were not included in the TARO estimation because their Census Day addresses were outside the test site and thus their data applies to other areas.

A procedure that might indicate the effect of including outmovers in the estimation would be to include the 409 inmovers as substitutes and impute match probabilities for them (since their match statuses are unknown). The treatments yielding the highest and lowest estimates in Table 2 have been applied to the TARO data with inmovers included; the results are displayed in Table 4. Note that the lower estimated undercount rates in Table 4 (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 2. 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 4 are larger than the corresponding estimates in Table 2 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 and other problem cases 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 optimal 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 for writing computer programs that processed the raw TARO data before imputation, and Robert O'Brien 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," Sociological Methodology 1987, 131-144.

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

FOOTNOTE

*This paper reports research undertaken by a member of the Census Bureau's staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

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

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 3

Parameter Estimates for the Additive Model (2) for Predicting the Estimated Undercount Rates in Table 2

	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 4

Estimated Undercount Rates (in %) by Race When Inmovers Are Included in the Data With Imputed Match Probabilities

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.