MISSING DATA IN THE 1990 POST ENUMERATION SURVEY

Philip M. Gbur, U.S. Bureau of the Census*
Statistical Support Division, Washington, D.C. 20233

Key Words: Imputation, Dual System Estimator, Census Coverage, Undercount

1. INTRODUCTION

The 1990 PES (Post-Enumeration Survey) was conducted after the 1990 Decennial Census to measure Census coverage error and to estimate adjustment factors which could be used to adjust the Census counts, should it be decided to make an adjustment. Various PES evaluation projects were designed to evaluate the 1990 PES results. This paper presents results from two of three evaluation projects to evaluate missing data.

The types of missing data which occur in the PES include nonresponse from whole household noninterviews, item nonresponse, and unresolved match statuses. Nonresponse from whole household noninterviews is compensated for by adjusting the weights of the interviewed households. Item nonresponse is eliminated by the use of a "hot-deck" imputation procedure and cases with unresolved match statuses have a match status probability imputed by means of a logistic regression model.

This paper provides a brief analysis of the missing data issues by examining three areas: 1) the percentage of noninterview and proxy interviews; 2) the information obtained from PES noninterviewed households converted to interview in the evaluations; and 3) the percent of imputation (hot-deck or logistic regression) for the PES. For each area, the effect of the missing data is examined, where possible.

2. OVERVIEW

2.1 PES Design and Procedures

Approximately 5,000 block clusters were selected for the PES. Census enumerations within these clusters comprise the PES E sample. The PES listed all housing units in these clusters independently of the census listings. These housing units were visited around July 1990 and interviews were obtained from the current occupants. The persons captured during this interviewing comprise the PES P sample.

Information could not be obtained from all households listed for the PES. Some household visits resulted in noninterviews. Minimal information was obtained from other households and the interview was classified as a "last resort" interview. In addition, some households were identified as whole household duplicates or whole household fictitious during PES processing. Each of these household types were treated as noninterview households. No information obtained from noninterview households was used in the PES processing and they were compensated for by a weight adjustment.

Even for interviewed households, certain information critical for PES processing may not be available for some persons. For missing person characteristics, a "hot-deck" imputation procedure assigned values for the missing characteristics based on the known characteristics of the persons. The PES P- and E-sample persons were processed through a matching operation to determine whether the P-sample persons matched and whether the E-sample persons were correctly enumerated. The information available during the matching operations may not be sufficient for a match or enumeration status to be determined. For persons with an unresolved status, a probability of match or correct enumeration was assigned based on a logistic regression model which considered the person's characteristics.

Further details on the PES and its design are in Woltman, Alberti, and Moriarity (1988) and Hogan (1990).

2.2 PES Evaluations Design and Procedures

The PES Evaluation sample is a stratified systematic subsample of about 920 PES sample block clusters. Within each of the block clusters, the Evaluation Followup (EFU) flagged selected PES P-sample cases including all nonmatched and unresolved persons. E-sample persons that had been sent to PES followup were sent out again as well.

Three questionnaires were used for the EFU including the PES Interview Questionnaire, PES Followup Questionnaire, and the PES Revisit Questionnaire (only the PES Interview Questionnaire and the PES Revisit Questionnaire were used for households selected for the evaluation projects discussed here). The PES Interview Questionnaire was used for households classified as noninterview following the initial PES interview. The Revisit Questionnaire, designed especially for the EFU operation, collected information for both P-sample and E-sample persons unresolved after PES matching.

A staff composed of only current survey interviewers was used for the EFU interviewing. The interviewers hired and trained for the PES and the Census were primarily temporary employees but often included current survey interviewers. For the EFU, the current survey interviewers were restricted to interview housing units that they had not interviewed during the production phases of PES.

Although interviewers were instructed to complete the EFU interview with a household member, the interviewers were permitted to obtain proxy information from a knowledgeable respondent (such as a landlord or neighbor). The results from the EFU interviewing were used in a matching operation by a team of matching experts using guidelines similar to those for PES.

2.3 Analytic Methodology

The PES Evaluation sample was designed to weight up to the PES national totals. In general, estimates which pertain to results of PES operations (such as percent noninterview) are presented unweighted while others (such as item imputation rates) are weighted.

The race variable examined in the analysis of percent imputation is a combination of race and hispanic origin.

Standard errors were calculated using a stratified jackknife variance estimator with VPLX - a variance estimation software system (Fay, 1990).

Differences examined in the analyses of variables for which standard errors are appropriate were tested at the 10 percent level of significance with t-tests.

3. RESULTS

3.1 Household Noninterviews

3.1.1 Outcome of Interview

The percent of eligible sample cases which resulted in a noninterview and the percent of interviews completed with a proxy respondent are given in Tables 1 and 2 for the PES pretests, PES, PES Followup (P-sample and E-sample), and PES Evaluations (P-sample, E-sample, and P-sample noninterview followup) samples. Note that a household classified as last resort, whole household duplicate, or whole household fictitious is considered a noninterview for the PES. However, proxy interviews are included as interviewed households; therefore, caution should be used when comparing these estimates with other surveys.

The 1990 PES has a higher percent noninterview and proxy than any of the PES pretests. (See Anolik (1988 and 1989), Childers and Hogan (1990), and Schenker (1988) for details on these PES pretests.) This is not unexpected since the 1990 PES was, by far, the largest of these surveys and survey operations become increasingly difficult to control with increasing size.

The percent of noninterviews is generally lower for the PES Followup than for the original PES interview while the percent for the PES Evaluations is generally higher than PES and PES Followup. This is reasonable as followup used generally higher quality interviewers but then repeated visits to a household may result in less cooperativeness in the later visits. Although the percent of noninterviews is high for the PES noninterview followup sample, it is lower than what may have been expected based upon results from studies conducted as part of the 1980 Post Enumeration Program and the 1985 Post Enumeration Survey (see Keeley (1984) and Childers (1986)). This may in part be due to acceptance of proxy interviews.

Interviews completed with a proxy respondent may obtain less complete or less accurate information than with a household member. The percent of proxy interviews is much higher for the PES Followup and the PES Evaluations than for the PES. However, the percent of proxy interviews in the PES Evaluations is markedly lower than for the PES Followup. This may result from the use of more experienced interviewers in the evaluations.

3.1.2 Examination of Converted PES Noninterview Households

The PES noninterview type by the evaluation outcome of interview for households treated as noninterviews is presented in Table 3. Overall, an unexpectedly high 83 percent of the households treated as noninterviews in the PES eligible for interviewing, are interviews for the evaluations. Of the PES refusal/not at home/other households eligible for interview in the evaluations, 75 percent are interviews.

3.1.3 Persons Within Converted Noninterview Households

Table 4 provides the percent of persons added in PES noninterview households interviewed in the evaluations by evaluation match status for PES noninterview type. Table 5 presents the estimated change in the number of matches resulting from interviewing households identified as noninterviews in the PES by evaluation poststratum. (As presented in Figure 1, the evaluation poststrata are defined by region, type of area, and whether the area has a high minority population or not.)

Of the persons added in PES noninterview households from an interview in the evaluations, 65 percent are matches. Of the persons added in households identified as whole household fictitious in the PES, 77 percent are matches. At the national level, an estimated 102,403 more matches than are indirectly added by the noninterview adjustment would be added to the PES from interviewing households identified as noninterviews in the PES. However, this estimate is not statistically significantly different from zero.

3.2 Imputation

3.2.1 Imputation Rates

The PES P-sample and E-sample percents of imputation are presented in Tables 6 and 7 for tenure, sex, age, race, and probability of match (P sample) or correct enumeration (E sample). The percent of imputation for characteristics in the PES E sample is significantly higher than in the P sample for all characteristics examined except tenure. Race and tenure have the highest percent of imputation for the P sample while race is markedly highest in the E sample. Although not shown, the imputation percents tend to be highest in the minority evaluation poststrata for the P and E samples and also in central city poststrata for the P sample.

3.2.2 Imputation and Estimated Census Coverage Error

Missing data may result in the incorrect assignment of match and enumeration status and the census undercount (overcount) estimates may then vary by the level of missing data. Correlations (Pearson correlations) between the percent of imputed characteristics and percent undercount were calculated based on evaluation poststrata level estimates and are presented in Table 8. The percent of imputation shows a relatively high level of association with the estimated Census undercount at the evaluation poststrata level, particularly for the E sample. The resulting high level of correlation may imply: 1) groups with a high undercount (overcount) rate are difficult to enumerate groups and may thus be expected to have high missing data rates and/or 2) high missing data rates may cause difficulties in assignment of match/enumeration status which result in high undercount (overcount) estimates. Lower observed correlations at the PES poststratum level (not shown), may reflect the large variability among the poststrata.

3.3. Evaluation of Match/Enumeration Status Imputation

In the PES, persons with an unresolved match status are assigned a probability of match. Unresolved cases in the evaluation sample clusters were sent to EFU and, as a result, many were resolved (assigned a match status of match or nonmatch).

The percent of PES P-sample persons with an unresolved match status in a match probability group by evaluation match status is given in Table 9. For example, of the PES unresolved persons which were assigned an imputed match probability of 0-25 percent, 6.38 percent were matched in the EFU and 43.82 percent were determined to be nonmatch. A higher percentage of persons have an unresolved match status, in the evaluations, than any other match status for all PES production match probability groups (0-25, 25-50, 50-75, and 75-100 percent), except the 0-25 percent group. The percentage of matches is lowest for cases with a match probability between 0 and 25 percent (although the percentage is not significantly different).

Comparing the probabilities from the production PES P sample with the PES evaluation results, the expected trend of higher match rates for cases with higher probabilities is supported. For persons

matched in EFU, their imputed probability of match is generally higher than for persons that were nonmatches in EFU. The highest observed unresolved rate is for the 75-100 percent group although persons with an extreme probability (high or low) would be expected to yield lower unresolved rates. Otherwise, for P-sample persons, the imputation process seems to exhibit results consistent with expectations.

Table 10 provides the percent of PES E-sample persons with an unresolved enumeration status in a correct enumeration probability group by evaluation enumeration status. The percent of cases resolved as correctly enumerated in the evaluations is the same for cases with an imputed correct enumeration probability of 0-50 percent as for those with a probability of 75-100 percent. A higher percent of cases are assigned an evaluation enumeration status of correctly enumerated than either of the other two enumeration status groups for all probability groups.

Comparing the probability of correct enumeration from PES with the EFU enumeration status, the trend of higher probability of correct enumeration with correctly enumerated EFU cases is lacking. The imputation model does not seem to behave as well as observed for the P sample. However, this may be distorted since the E-sample tabulations do not account for unresolved geocodes nor duplicates in surrounding blocks.

4. SUMMARY

Not unexpectedly, the 1990 PES has a higher percent noninterview and proxy than any of the PES pretests. Also, the percent of noninterviews is generally higher for the original PES interview than for the PES followup, while the percent for the PES Evaluations is generally higher than PES and PES Followup. This is reasonable as followup used generally higher quality interviewers but repeated visits to a household may result in less cooperativeness in the later visits. However, the percent of proxy interviews is much lower for the PES than for the PES Evaluations which is still lower than the PES Followup.

Overall, the Evaluations were generally successful in converting PES noninterviews. Specifically, 83 percent of the households treated as noninterviews in the PES eligible for interviewing, are interviews for the evaluations. Of the persons added from these households, 65 percent are matches. Based on these matches, at the national level, (although not significantly different from zero) an estimated 102,403 more matches than are indirectly added by the noninterview adjustment would be added to the PES.

The percent of imputation for characteristics in the PES E sample is significantly higher than in the P sample for all characteristics examined except tenure. Race and tenure have the highest percent of imputation for the P sample while race is markedly highest in the E sample. The imputation percents tend to be highest in the minority evaluation poststrata for the P and E samples and also in central city poststrata for the P sample.

Missing data may result in the incorrect assignment of match and enumeration status and the census undercount (overcount) estimates may then vary by the level of missing data. The estimated correlations between the percent of imputed characteristics and percent undercount show a relatively high level of association at the evaluation poststrata level, particularly for the E sample. This result may imply that difficult to enumerate groups should be expected to have high missing data rates and/or high missing data rates contribute to high undercount estimates.

Comparing the probabilities of match from the production PES P sample imputation algorithm with the PES evaluation results, the expected trend of higher match rates for cases with higher probabilities is supported. The highest observed unresolved rate is for the 75-100 percent group although persons with an extreme probability (high or low) would be expected to yield lower unresolved rates. Otherwise, for P-sample persons, the imputation process seems to exhibit results consistent with expectations. Comparing the probability of correct enumeration from the PES E sample imputation algorithm with the EFU enumeration status, the trend of higher probability of correct enumeration with correctly enumerated EFU cases is lacking. The imputation model does not seem to behave as well as observed for the P sample. However, this may be distorted since the E-sample tabulations do not account for unresolved geocodes nor duplicates in surrounding blocks.

REFERENCES

Anolik, Irwin, (1988), "The 1986 Rural Post-Enumeration Survey in East Central Mississippi", Proceedings of the Survey Research Section of the American Statistical Association, pp. 576-581.

Anolik, Irwin, (1989), "The 1987 Post-Enumeration Survey", Proceedings of the Survey Research Section of the American Statistical Association, pp. 710-715.

Childers, Danny R., (1986), "The 1985 Post Enumeration Survey", 1985 Test Census Preliminary Research and Evaluation Memorandum No. 63, internal Census Bureau memorandum.

Childers, Danny R., and H. Hogan, (1990), "Results of the 1988 Dress Rehearsal Post Enumeration Survey", Proceedings of the Survey Research Section of the American Statistical Association, pp. 547-552.

Fay, Robert E., (1990), "VPLX: Variance Estimates for Complex Samples", Proceedings of the Survey Research Section of the American Statistical Association, pp. 266-271.

Hogan, Howard, (1990), "The 1990 Post-Enumeration Survey: An Overview", Proceedings of the Survey Research Section of the American Statistical Association, pp. 518-523.

Keeley, Catherine, (1984), "The Post Enumeration Program Unresolved Cases Study Pretest", Preliminary Evaluation and Results Memorandum No. 74, internal Census Bureau memorandum.

Schenker, Nathaniel, (1988), "Handling Missing Data in Coverage Estimation, with Application to the 1986 Test of Adjustment Related Operations", Survey Methodology, 14, No. 1, pp 87-97.

Woltman, H., N. Alberti, and C. Moriarity, (1988), "Sample Design for the 1990 Census Post Enumeration Survey", Proceedings of the Survey Research Section of the American Statistical Association, pp. 529-534.

ACKNOWLEDGEMENTS

The author wishes to thank Mary Mulry for suggesting and encouraging this paper; and Jon Clark and David Bateman for guidance throughout the project and comments on the paper. In addition, without the dedicated work of innumerable Bureau personnel who contributed to the PES and PES Evaluations, this paper would not have been possible.

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

FIGURE 1: EVALUATION POSTSTRATA

1 Northeast - Central City - Minority
2 Northeast - Central City - Nonminority
3 U.S. - Noncentral City - Minority
4 Northeast - Noncentral City - Nonminority
5 South - Central City - Minority
6 South - Central City - Nonminority
7 South - Noncentral City - Nonminority
8 Midwest - Central City - Minority
9 Midwest - Central City - Nonminority
10 Midwest - Noncentral City - Nonminority
11 West - Central City - Minority
12 West - Central City - Nonminority
13 West - Noncentral City - Nonminority

TABLE 1: PERCENT NONINTERVIEW AND PROXY INTERVIEW FOR PES PRETESTS AND THE 1990 PES

Study	Occupied Units	Percent Noninterview	Interviewed Units	Percent Proxy
1990 PES	143913	1.56	141667	4.25
1986 Rural PES	3252	0.00	2910	1.92
1988 Dress Rehearsal PES				
- St. Louis/Columbia	8584	1.13	8487	3.35
- Washington	960	0.21	958	3.55
1987 PES	1446	0.14	1444	3.67
1986 TARO	5935	0.54	5903	3.20

TABLE 2: PERCENT NONINTERVIEW AND PROXY FOR PES, PES FOLLOWUP, AND PES EVALUATION SAMPLES

Sample	Occupied Units	Percent Noninterview	Interviewed Units	Percent Proxy
PES	143913	1.56	141667	4.25
PES FU				
P-sample	17936	1.37	17690	17.28
E-sample	29402	1.28	29029	19.80
PES Eval				
P-sample	5057	1.88	4962	10.62
E-sample	5506	1.36	5431	13.63
NI FU	212	16.98	176	13.64

TABLE 3: PES NONINTERVIEW TYPE BY EVALUATION OUTCOME OF INTERVIEW

PES Noninterview Type	Evaluation Outcome of Interview			
	Total	Interview	Noninterview	Out of Scope
Total	257	176	36	45
Refusal/Not at Home/Other	146	89	30	27
Whole Household Duplicate	71	53	4	14
Whole Household Fictitious	40	34	2	4

TABLE 4: PES NONINTERVIEW TYPE BY EVALUATION MATCH STATUS FOR PERSONS ADDED IN INTERVIEWED PES NONINTERVIEWS

PES Noninterview Type	Evaluation Match Status (%)				
	Total	Match	Nonmatch	Unresolved	Out of Scope
Total	443	64.79	18.96	8.35	7.90
Refusal/Not at Home/Other	250	68.00	18.00	10.80	3.20
Whole Household Duplicate	95	44.21	24.21	3.16	28.42
Whole Household Fictitious	98	76.53	16.33	7.14	0.00

TABLE 5: ESTIMATED CHANGE IN NUMBER OF MATCHES FROM PES NONINTERVIEWS INTERVIEWED IN THE EVALUATIONS BY EVALUATION POSTSTRATUM

Evaluation Poststratum	Change	Standard Error
Total	102403	82165
1	24728	15860
2	13278	12986
3	-3815	9552
4	10417	9855
5	937	4700
6	-1613	1065
7	14630	60336
8	373	3324
9	10800	9385
10	22187	26804
11	19917	19604
12	8235	10236
13	-17672	30285

TABLE 6: PERCENT IMPUTATION FOR PES P SAMPLE AND E SAMPLE BY CHARACTERISTIC

Estimate	Tenure	Sex	Age	Race
P sample Base = 240,651,222				
Percent	2.26	0.51	0.71	2.49
SE	0.11	0.04	0.05	0.11
E sample Base = 244,200,930				
Percent	2.48	1.04	2.39	11.75
SE	0.08	0.04	0.13	0.24

SE = Standard Error

TABLE 7: PERCENT IMPUTATION FOR PES P-SAMPLE PROBABILITY OF MATCH AND E-SAMPLE PROBABILITY OF CORRECT ENUMERATION

Estimate	P-sample Probability of Match	E-sample Probability of Correct Enumeration
Base	240,651,222	244,200,930
Percent	1.70	2.11
Standard Error	0.06	0.12

TABLE 8: CORRELATIONS BETWEEN THE PERCENT OF IMPUTED CHARACTERISTICS AND PERCENT UNDERCOUNT FOR SELECTED CHARACTERISTICS BY THE PES P SAMPLE AND E SAMPLE

Characteristic	P Sample	E Sample
Tenure	0.49 (0.09)	0.80 (0.00)
Sex	0.52 (0.07)	0.84 (0.00)
Age	0.56 (0.05)	0.72 (0.01)
Race	0.73 (0.00)	0.82 (0.00)
Probability of Match (P sample) or Correct Enumeration (E sample)	0.71 (0.01)	0.77 (0.00)

(X.XX) = p-value for testing H_o: rho=0 (all numbers rounded to two decimal places)

TABLE 9: PERCENT OF P-SAMPLE UNRESOLVED PERSONS FOR EVALUATION MATCH STATUS BY PES IMPUTED MATCH PROBABILITY

Evaluation Match Status	Imputed Match Probability (%)				
	Total	0-25	25-50	50-75	75-100
Base (Weighted)	2771717	1065958	571759	659732	474268
Match	11.64 (2.05)	6.38 (2.15)	12.85 (4.31)	16.54 (4.77)	15.21 (5.29)
Nonmatch	27.03 (4.50)	43.82 (7.15)	26.10 (7.03)	15.38 (3.83)	6.63 (3.89)
Unresolved	58.55 (4.38)	46.65 (6.56)	56.71 (7.06)	65.94 (5.96)	77.25 (6.30)
Out of Scope	2.77 (0.91)	3.15 (1.57)	4.34 (2.90)	2.14 (1.47)	0.91 (0.86)

(X.XX) = Standard Error

TABLE 10: PERCENT OF E-SAMPLE UNRESOLVED PERSONS FOR EVALUATION ENUMERATION STATUS BY PES IMPUTED PROBABILITY OF CORRECT ENUMERATION

Evaluation Enumeration Status	Imputed Probability of Correct Enumeration (%)			
	Total	0-50	50-75	75-100
Base (Weighted)	2097296	223892	461252	1412151
Correct Enumeration	62.19 (5.04)	67.07 (10.25)	50.18 (10.03)	65.34 (5.96)
Erroneous Enumeration	16.96 (3.39)	10.15 (3.72)	25.88 (8.22)	15.13 (4.17)
Unresolved	20.85 (3.71)	22.78 (8.88)	23.94 (6.80)	19.54 (4.18)

(X.XX) = Standard Error