HANDLING UNMATCHABLE CASES IN THE POST-ENUMERATION SURVEY

Gregg Diffendal, U.S. Bureau of the Census*
Washington, D.C. 20233

Key Words: Census undercount, dual-system estimation

Wolter (1986) provided the underlying assumptions for the application of capture-recapture models for human populations. These assumptions were essentially used in the 1990 Post-Enumeration Survey (PES). One area of difference between the assumptions listed in Wolter (1986) and the 1990 PES is the handling of "unmatchable" cases in the census. Wolter assumes that all census cases are matchable or are spurious events. Spurious events are census (or P sample) inclusions that should be eliminated prior to estimation. This paper examines the unmatchable cases in the E sample and describes the underlying assumptions for these cases as used in the 1990 PES. This formulation has implications for examining the 2 x 2 table and for combining with demographic analysis. In addition, alternate assumption for the unmatchable cases are used to produce an alternate missing data model for the 1990 PES.

Heuristically, the PES assigns every person to one of the four cells given in Table 1.

Table 1

		P-Sample In	P-Sample Out	
Census	In	x_{11}	x_{12}	x_{1+}
	Out	x_{21}	x_{22}	
		x_{+1}		

Where x_{11} are persons included in the Census and in the P sample, x_{21} are persons included in the P sample and missed in the census, x_{12} are persons included in the Census and missed in the P sample and x_{22} are persons missed in the P sample and in the Census. The + refers to summation over a subscript. x_{22} is unobserved and is typically estimated assuming independence between the P sample and the census and is estimated by

$\hat{x}_{22} = \frac{x_{12}x_{21}}{x_{11}}$. The value x_{1+} is not the Census count. Instead the E sample of the PES is used to measure erroneous inclusion (spurious events) and unmatchable records which are subtracted from the Census count. In addition, census whole person substitutions (imputations) are subtracted from the census count to derive the x_{1+} used in Table 1.

This paper examines the group of cases that are being treated as being out of the census because of insufficient information for matching. The corresponding P-sample cases that potentially could be matched to these cases are called nonmatches. This inflates the in-P sample, out-of-Census and out-of-P sample, out-of-Census categories and deflates the in-Census, in-PES and in-Census out-of-Census categories. Suitable modifications to the 2 x 2 table should give better interpretation of the individual cells.

The E sample is described first focusing on the measurement of unmatchable persons. The second section provides a means of handling unmatchable case by expanding the 2 x 2 table. The list of assumptions given by Wolter (1986) are modified to incorporate unmatchable cases. Finally, a few implications of this work are given with examples from the 1990 PES.

E Sample Measurement

The E sample is a sample from the census used to estimate the number of erroneous inclusions and unmatchable records. A list of cases that are treated as erroneous (and unmatchable) and subtracted from the census count for the dual system estimator are: duplicate enumerations, fictitious enumerations, geocoding errors, records with no name, general erroneous enumeration (e.g. born after Census Day, died before Census Day, should be counted at another address) and a proportion of unresolved cases. Census whole person imputations (substitutions) are also treated as erroneous (subtracted from the census count), but are measured directly in the census and so are not estimated by the E sample (but could be if desired).

Although all of these cases are treated as erroneous enumeration, many represent persons correctly counted in the census. They are treated in this manner since they cannot be called matched or not matched to the P sample with certainty. Therefore, all erroneous enumerations are not spurious events. Some erroneous enumerations are unmatchable and treated as being out of the census. Unmatchable cases could be called matches if more complete information was obtained. An example of an unmatchable case is a census case without a recorded name (coded as a "K"). Other E-sample cases may be viewed as including some unmatchable cases. Fictitious enumeration can be viewed as a field imputation for the household. Some duplicate enumerations occur when an enumerated household is enumerated again instead of visiting the unenumerated household. Census substitutions may also represent persons captured in the census, but the enumerator was unable to obtain responses to the census questionnaire.

Some types of erroneous enumerations are all spurious events in the census. College students counted at home and at college should only be included once. A person born after census day is also a spurious event. The next section distinguishes between spurious events from the unmatchable cases by modifying the assumption given in Wolter (1986).

Assumptions Needed for Unmatchable Case

All of the assumptions for the Petersen model as given in Wolter (1986) are restated for completeness and since some changes are needed for many of them.

1. The Closure Assumption. The population V is closed and of fixed size N. This is the same assumption as given in Wolter.

2. The Multinomial Assumption. Let E_i denote the

multinomial distribution with the following parameters:

List A		List B: In	List B: Out	
In (Matchable)		P_{i11}	P_{i12}	P_{i1+}
Out		P_{i21}	P_{i22}	P_{i2+}
In (Unmatchable)		P_{i31}	P_{i32}	P_{i3+}
		P_{i+1}	P_{i+2}	1

The event that the i_{th} individual is in (matchable), out, or in (unmatchable) in List A and in or out List B is correctly modeled by the distribution E_i.

Note that an additional row for the unmatchable cases has been added to List A, the departure from Wolter to handle unmatchable persons.

3. Autonomous Independence. List A and List B are created as a result of N mutually independent trials, using distribution $E_1, E_2, \ldots E_N$. The resulting data are:

List A		List B: In	List B: Out	
In (Matchable)		x_{11}	x_{12}	x_{1+}
Out		x_{21}	x_{22}	x_{2+}
In (Unmatchable)		x_{31}	x_{32}	x_{3+}
		x_{+1}	x_{+2}	$x_{++} = N.$

where $x_{ab} = \sum x_{iab}$ and x_{iab} is an indicator random variable signifying whether the i_{th} individual is in cell (a, b) for a = 1, 2, 3, + and b = 1, 2, +. The census count minus erroneous enumerations, substitutions and unmatchable is x_{1+} and is considered observable. The cell count x_{11}, x_{12}, $(x_{21} + x_{31})$ are considered observable on the basis of the survey data and subsequent matching to the census. Also x_{3+} is considered observable.

Note that the usually 2 x 2 table uses $x_{21}+x_{31}$ and $x_{22}+x_{32}$ for the values for x_{21} and x_{22} in Table 1 since the In (unmatchable) cases are treated as being out of the census.

4. The Matching Assumption. It is possible to assign every individual recorded in the sample as being In (matchable), Out or In (unmatchable) to the Census. That is, every individual can be assigned to a cell in the 2 x 3 table.

5. Spurious Events Assumptions. This is the same as stated in Wolter (1986), that both lists A and B are void of spurious events or they are eliminated prior to estimation. Clearly spurious events do occur in the P sample and in the census and are taken out of each count. Note that for the census the estimated number of erroneous enumerations and census substitutions are reduced by our estimate for x_{3+}.

6. The Nonresponse Assumption. This is the same as stated by Wolter (1986). Sufficient identifying information is obtained about the nonrespondents in both the census and the sample survey to permit exact matching. Some degree of nonresponse will exit in the census and the sample survey.

7. The Poststratification Assumption. The same as stated in Wolter (1986). Note that this assumption is critical since many In (unmatchable) cases often have many missing characteristics that are imputed before being used in poststratification.

8. Causal Independence. This assumption is expanded to include the In (unmatchables) to be independent of List B. That is, the event of being In (matchable) in List A is independent of being included in List B and the event of being In (unmatchable) in List A is independent of being included in List B. The cross product ratio satisfy

$$P_{i11}\, P_{i22} /(P_{i12}\, P_{i21}) = P_{i11}\, P_{i32}/(P_{i12}\, P_{i31}) = 1$$

9. (Wolters assumption number 11). For the Petersen model, we assume

$$P_{i1+} = P_{1+}, \; P_{i+1} = P_{+1}.$$

Implications

Adding the In (unmatchable) cases further develops the underlying models for the PES. Without an estimate of x_{3+}, the unmatchable cases in the Census, this work would not have practical uses. Fortunately, the coding of the E sample allows us to estimate the In (unmatchables) in the census. The two extreme estimates are that all erroneous enumerations (EE) and census substitutions (II) are In (unmatchable) or that all EE's and II's are spurious events. Clearly census cases without names ("K") could be treated as In (unmatchables). Also, census substitutions (II) could be treated as In (unmatchables). For the rest of the paper, I will use census cases without names as the estimated number of In (unmatchables).

This work may be applicable to combining the PES and demographic analysis. The 2 X 2 table that has been used for combining which I shall call Model A is

	PES: In	PES: Out
In (Matchable)	x_{11}	x_{12}
Census Out+ In (Unmatchable)	x_{21+} x_{31}	x_{22+} x_{32}

but I believe we should be using

		PES: In	PES: Out
Census	In	x_{11+} x_{31}	x_{12+} x_{32}
	Out	x_{21}	x_{22}

which I shall call Model B. The estimate for x_{31} and

x_{32} are given by

$$\hat{x}_{31} = \frac{x_{11}}{x_{1+}} x_{3+} \text{ and } \hat{x}_{32} = \frac{x_{12}}{x_{1+}} x_{3+}. \quad (1)$$

The net effect is to lower the estimated x_{22} cell by our estimate of x_{32}. Note that the estimated $x_{22} + x_{12}$ cells for the 2 X 2 table will give the appearance of adding more people from the independence assumption as compared to using Model B. The examples in the next section will make this clear. For combining demographic analysis and the PES, demographic analysis is used to create a measure of association in the 2 x 2 table (Bell 1991) at the national level. The PES has separate estimates at sub-national levels. Differences in the number of the In (unmatchable) cases at the subnational level may cause inaccuracies even if the other underlying assumptions are roughly correct.

Examples

Table 2 and 3 provide the data from the 1990 PES using 357 Poststrata design (Hogan 1992). The totals reported are the estimates assuming independence, not the estimates summed over each poststrata. The data in Table 2 are for black males age 18-29 in the U.S. and Table 3 is for nonblack females 18-29 in the U.S.

Table 2a is the usual 2 X 2 table for black males age 18-29. Note that the estimated coverage of the population indicates the P sample was slightly better than the correctly enumerated Census coverage. Table 2b provides the breakouts of the 2 X 3 table using census cases without names ("K") as our estimate of In (unmatchables). Table 2c provides the data as stated in model B. When using model B, the coverage of the census and the P sample are comparable. The x_{22} cell for Model B is over 10% lower than the x_{22} cell for Model A. Similar observations were observed for the other age-sex categories for the Black population.

Table 3a, 3b, and 3c correspond to Table 2a, 2b and 2c, but for nonblack females age 18-29. Again Model A shows that the PES has better coverage than the census. Model B in table 3c does not substantially change this observation as it did for the black males 18-29. This also held for the other nonblack categories except for females age 50+. The individual cell estimates for in-census out- of-P sample (x_{12}) were negative for nonblacks age 30 - 49 males and females. In this case the interpretation of the individual cells is not clear. Note again that over 10% of x_{22} cells under Model A is explained by the In (unmatchables).

Alternate Missing Data Model

In the E sample, cases without names ("K") were treated as being out of the census. In the P sample, cases without names ("J") were treated as unresolved and were imputed using a hierarchical logistic regression model (Belin...Zaslavsky 1992). The P-sample nonmovers without names were imputed with a mean probability of being matched at .93. This is slightly higher than the .92 average match rate for all resolved cases.

Let's assume that a respondent who does not give his name in the census is also not going to give his name in the P-sample interview. In this case the "J" case in the P sample should be considered a nonmatch rather than the imputed value which essentially calls the case a match. So the usual values given in Model A, the imputation model is creating too many-in PES, in-Census (x_{11}) without a suitable increase in the in Census total (x_{1+}). This may be a partial explanation of the results noted by Bell (1991) that around 30% of the 2 x 2 tables had the estimated x_{11} cell larger than the corrected Census marginal (x_{1+}).

The count of the number of J's with a corresponding K in the same block cluster is 681 out of a total of 1766 J's in the P sample or 38.6 of the time. Therefore, the predicted probability of the J's are reduced by .386 under this assumption.

Under this alternate missing data model, the resulting dual system estimate for Black males age 18 - 29 is 2,826,600, a 0.26% increase in the DSE. For Nonblack females age 18 - 29, the DSE is 20,774,491, a 0.11% increase.

References

Bell, W.R. (1991) "Using Information from Demographic Analysis in Post-Enumeration Survey (PES) Estimation - New Methods and Further Results" paper presented at the 1991 American Statistical Association Meeting.

Wolter, K.M. (1986) "Some Coverage Error Models for Census Data," Journal of the American Statistical Association, 81, 338-353.

Wolter, K.M. (1990) "Capture-Recapture Estimation in the Presence of a Known Sex Ratio", Biometrics, 46, 157-162.

Hogan, H. (1990) "The 1990 PES: An Overview", Proceedings of the American Statistical Association, Survey Research Methods Section.

Belin, T.R., Diffendal, G.J., Mack, S., Rubin, D.B. Schafer, J.S., Zaslavsky, A.M., (1992) "Hierarchical Logistic-Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation", paper submitted for publication.

Hogan, H. (1992) "The 1990 Post-Enumeration Survey: Operations and New Estimates", paper to be presented at the American Statistical Association 1992.

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

Table 2

Black Males Age 18-29
P-Sample

		In	Out	
Census	In (Matchable)	1911911	368567	2280478
	Out + In (Unmatchable)	451577	87053	
		2363488		2819108

		In	Out	
Census	In(Matchable)	1911911	368567	2280478
	Out	377532	72779	
	In (Unmatchable)	74045	14274	88319

		In	Out	
Census	In (Matchable + Unmatchables)	1985956	382841	2368797
	Out	377532	72779	
		2363488		2819108

Table 3

Nonblack Females Age 18 - 29
P-Sample

		In	Out	
Census	In (Matchable)	17816476	675934	18492410
	Out + In (Unmatchable)	2176342	82568	
		19992818		20751320

		In	Out	
Census	In (Matchable)	17816476	675934	18492410
	Out	1906100	72315	
	In (Unmatchable)	270242	10253	280495
		19992818		20751320

		In	Out	
Census	In (Matchable + Unmatchable)	18086718	686187	18772906
	Out	1906100	72315	
		19992818		20751320