AN INVESTIGATION OF LATENT CLASS MODELS FOR EVALUATING CENSUS COVERAGE ERROR

Paul Biemer, Research Triangle Institute, James Treat, Henry Woltman, & Ann Vacca, Bureau of the Census
Paul Biemer, RTI, P.O. Box 12194, Research Triangle Park, NC 27709

Key Words: Reinterview, Undercount, Measurement Bias

1. Introduction

The reinterview survey is an important method for estimating and reducing nonsampling errors in surveys, particularly reinterview surveys that seek the truth - so-called *true-value* or *gold standard* reinterviews. In these surveys, a sample of survey elements are reinterviewed to measure the same characteristics obtained in the first interview. This may require altering the wording of the question so that the time reference periods correspond identically. Other modifications may be necessary to ensure that any differences between the interview and reinterview responses are due to measurement error and not changes in the characteristics being measured. Once these discrepancies between the interview and the reinterview survey results are identified, their differences are reconciled; that is, the discrepancies between the first and second responses are discussed with the respondent for the purpose of arriving at the "best" response. The reconciled measurement is then assumed to be the truth for purposes of evaluating the measurement bias in the original responses.

The Census Bureau uses this reconciled reinterview technique in a number of ongoing, national surveys (for example, the National Crime Survey and the Survey of Income and Program Participation) and is planning to employ a version of it in the next Decennial Census to obtain a more accurate count of the number of persons in each household. The current paper focuses on a method for evaluating the quality of data collected in these types of reinterview surveys using latent class models. Traditionally, the accuracy of data from true-value reinterviews was assessed by conducting yet another true-value reinterview survey of the reinterview respondents using the best personnel and the best methods affordable. In this "reinterview of the reinterview," reconciliation methods are again employed to obtain a second, reconciled response for each item to be evaluated. This third measurement is then regarded the gold standard since it is (presumably) the best measurement available. The gold standard measurements would then be used to evaluate not only the reinterview measurements but also the original survey measurements.

In this paper we explore the potential of latent class models for evaluating the error in all three measurements: the original survey, the reinterview, and the reinterview of the reinterview. These models and methods do *not* require the assumption that any of the measurements are infallible and thus are well suited for the purposes of quality evaluation. Although latent class models can be applied to the situation of two measurements - say, an interview and a reinterview, with three measurements the additional degrees of freedom allow for greater numbers of parameters to be fit and thus better fitting models. Further, the additional degrees of freedom also provide opportunities to test model lack-of-fit. With only two measurements, all the degrees of freedom are typically used up by the fitted parameters.

To illustrate the concepts, we use data from an evaluation of the 1995 Census Test: a pretest conducted by the Census Bureau of procedures planned for the Census 2000. The 1995 Census Test evaluation provided three measurements of the key Census characteristic, viz., persons living in the household on Census Day. The three measurements are: a) the respondent's mail or interviewer-assisted response to the Census questionnaire regarding who lives in the household; b) the respondent's response to an interviewer administered reinterview regarding the same; and c) the respondent's response to a second reinterview using methods which were substantially improved over those in (a) and (b).

In the next section, we describe the 1995 Census Test evaluation study and the data set that we are working with in this demonstration of latent class models. Section 3 provides a summary of the notation and assumptions of the latent class models approach. Section 4 discusses the models that we will consider in our analysis of the Census Test evaluation data. Section 5 illustrates the use of these models and methods for our data set and, in Section 6, we conclude with a discussion of the major lessons learned from our investigation.

2. The 1995 Census Test Evaluation

Background

The objective of the 1995 Census Test evaluation was to measure and evaluate the quality of the Integrated Coverage Measurement (ICM) Person Interview data. The evaluation focused on errors that are relevant to the study of census coverage estimator bias. The data for the evaluation were obtained from the 1995 ICM Evaluation Interview survey.

Different types of nonsampling errors might be introduced in the data during collection and processing. Data collection errors can be caused by the instrument design, the enumerators or the respondents. In the ICM Person Interview, examples of data collection error include respondent recall errors on Census Day residents or misinterpretations of Census residency rules. Errors caused by an enumerator would include falsifying part of a roster (e.g., fabricating people) or entering erroneous information during the reconcilation of differences between Census and ICM rosters. Finally, data processing errors might be introduced when Census and ICM Person Interview data are linked and residency status is established. To determine the overall accuracy of the Person Interview results, a reinterview was conducted on a sample of the ICM households. This reinterview was designed to obtain the "best" residency status classifications for all sample persons as described below.

ICM Person Interview

The ICM Person Interview was a computer assisted personal interview (CAPI). The purpose of the interview was to obtain an independent roster of names and demographic information for each person living at a sample address on Census Day. After collecting the independent roster and confirming it with the respondent, the enumerator was permitted access to the Census roster. The Census roster was completed either by the respondent alone on a self-administered form or by the an enumerator interviewing the respondent during the nonresponse followup operations. The Census roster information had been keyed and loaded into the instrument before the interview.

The Census roster and the independent roster were compared, first by computer software using an exact match on age and sex, and second by the enumerator visually examining the data. Persons that appeared on both rosters were linked and considered to be correctly enumerated. For the persons *not* linked on the ICM independent roster (ICM nonmatch persons) the respondents were guided through a separate instrument path. On this path, information was gathered on the reasons why these persons were listed on the ICM independent roster, but not on the Census roster. Data were obtained on the living situation of these persons on Census Day to determine their residence status.

Similarly, for the persons not linked on the Census roster (Census nonmatch persons) the respondents were guided through a path to obtain information on the reasons why these persons were listed on the Census roster, but not on the independent roster. Data were obtained on the living situation of these persons on Census Day to determine their residence status.

Evaluation Interview

The Evaluation Interview was a CAPI reinterview for a subsample of the ICM households that was modeled after the ICM Person Interview. The main difference between the two interviews were the data collection and processing staff. The Evaluation Interview employed the most qualified and skilled enumerators available. Similarly, the best staff in the processing office conducted the editing and coding associated with the Evaluation Interview.

Like the ICM Person Interview, the Evaluation Interview also collected an independent roster of names and demographic information for each person living in the household on Census Day. This roster was confirmed for accuracy with the respondent and then the enumerator was permitted access to the Evaluation Interview input roster. This input roster was the union of persons from the initial Census roster plus any ICM nonmatched persons.

The two rosters were compared, first by computer matching on age and sex and second by the enumerator visually inspecting the data. Persons listed on both rosters were linked. For persons who were not linked on the Evaluation Interview independent roster (Evaluation Interview nonmatched persons) the interview was guided through a separate path of the instrument to find out why these persons were listed on the Evaluation Interview independent roster, but not on the input roster. For the persons not linked on the input roster (input roster nonmatched persons) the interview took a different instrument path, again to establish why these persons were listed on the input roster, but not on the Evaluation Interview independent roster, and to determine their true residence status.

Similar to the ICM Person Interview data process, Evaluation Interview cases with unresolved residence status and/or enumerator notes attached were sent to a clerical review operation in the processing office. The purpose of this review was to determine the correct residence status of these cases.

Data Collection

The Evaluation Interview was conducted in Oakland, California. In this site, the ICM sample consisted of two subsamples with approximately 5,000 households in each. The Evaluation Interview sample was selected from the panel that would not be receiving a followup visit for an ICM production operation was being conducted simultaneously with this evaluation. The ICM Person Interview field work began on June 5, 1995. The sample selection was based on data collected in the ICM Person Interview by July 17, 1995. The cut-off date was required in order to have the input roster loaded into the instrument and to have the enumerator assignments ready before the start of the reinterview on August 7, 1995. Only cases with outcome codes of "completed" and "partially" completed were eligible for sampling. Thus, the sampling universe was reduced to about 3,400 housing units, or 68 percent of the original sample. approximately 1,000 housing were selected from this universe. A number of cases that had previously received a quality control reinterview were eliminated from the sample, leaving a total of 948 housing units to be interviewed.

The sample design called for a stratified cluster sample. The housing units were stratified by number of Census and independent roster nonmatches and by ICM Person Interview outcome codes. The strata definitions and sample sizes (n) in housing units are presented below:

Stratum 1: Whole household match. All Census and independent roster persons match (n=116)

Stratum 2: At least one person match between the Census and the independent roster, at least one Census roster nonmatch and no independent roster nonmatches (n=153)

Stratum 3: At least one person match between the Census and the independent rosters, at least one independent roster nonmatch and no Census roster nonmatches (n=195)

Stratum 4: At least one person match between the Census and the independent rosters, at least one Census nonmatch and at least one independent roster nonmatch (n=83)

Stratum 5: Whole household nonmatch with zero persons on the Census roster (n=238)

Stratum 6: Whole household nonmatch with at least one person on the Census roster (n=162)

Since most rostering errors are expected to come from strata 2, 3, 4, and 6, i.e., the unit variability in these strata are expected to be greater than in strata 1 and 5, the former strata were oversampled to reduce the variance of an estimator based on the number of errors.

ICM methodology was concerned with estimating the coverage error associated with the initial Census responses. ICM was used to produce an estimate of the missed or erroneously included persons in the initial Census. The Evaluation Interview was designed to provide a picture of the accuracy of the ICM data collection and processing. It was used by the Census Bureau to examine factors related to the bias of the ICM coverage estimator. (See West and Griffiths, 1996 for a full report of the Evaluation Interview results.)

In what follows, we re-analyze the ICM and Evaluation Interview data using a latent class modeling approach. The advantage of our approach is that, unlike the analysis conducted by West and Griffiths in their report, it is not necessary to assume that the Evaluation Interview results are the gold standard. Our approach estimates the error rates associated with all three classification systems: Census, ICM, and Evaluation Interview. In addition, it provides an estimate of the total number of persons in the target population based upon the information collected in all three systems.

3. Description of the Three Classifiers

Let the universe of potential population members be donated by U and denote by $P \subseteq U$, the subset of U that is truly in the population. Thus, the set $U \sim P$ is composed of persons who are not in the population, fabrications created by respondents and enumerators, and other rostering (enumeration) errors. For each element $i \in U$, define

$$X_i = \begin{cases} 1 & \text{if } i \in P \\ 2 & \text{if } i \notin P \end{cases} .$$

Thus, X_i is an indicator variable for membership in P. In what follows, we consider three processes for classifying members of U: the household Census classification (referred to as the Census), the Integrated Coverage Measurement (ICM) reinterview of Census respondents, and the Evaluation Interview.

Census Classification

As described above, the Census classification process is conducted by a household respondent or enumerator working with the respondent. The respondent considers the potential members of the population and, using the census instructions as a guide, either lists the potential member on the roster, and thereby classifies the person as in P, or does not list the potential member classifying him/her as not in P.

For $i \in U$, define

$$A_i = \begin{cases} 1 & \text{if listed on the Census roster} \\ 2 & \text{if not listed} \end{cases} .$$

ICM Classification

The ICM interviews are conducted by enumerators using Computer Assisted Personal Interviewing (CAPI). First, an independent Census Day roster is obtained from the respondent. Then this roster is compared with the Census roster for the household and any differences are reconciled. The result of this reconciliation process is a classification of each person on the combined Census and ICM roster.

For $i \in U$, let B_i denote the ICM classification which has three states as follows:

$$B_i = \begin{cases} 1 & \text{if classified as in } P \\ 2 & \text{if classified as not in } P \\ 3 & \text{if classified as unresolved or status unknown} \end{cases} .$$

When the ICM enumerator cannot decide whether a person should be classified as a resident, he/she may classify the individual as "unresolved" or status unknown. Subsequently, these unresolved cases are classified as residents or non-residents by coders or these classifications are imputed using a computer model.

Evaluation Interview Classification

The Evaluation Interview is a second reinterview of the census respondents using procedures similar to the ICM interview. In the Evaluation Interview, a third roster of the Census Day residents is constructed by the household respondent using free recall and without referring to the previous two rosters. Then Evaluation Interview roster is compared to the combined Census and ICM roster and any differences are reconciled with the respondent. The result of this reconciliation process is the Evaluation Interview classification, C_i, defined for $i \in U$ as:

$$C_i = \begin{cases} 1 & \text{if classified as in } P \\ 2 & \text{if classified as not in } P \\ 3 & \text{if classified as unresolved or status unknown} \end{cases}$$

Unresolved cases are handled in the same manner as described for the ICM.

4. Model Assumptions and Notation

Latent Class Model

For any two variables E_i and F_i defined for $i \in U$, let $\pi_{ef(i)}$ denote $P(E_i = e, F_i = f)$. The conditional probability $P(E_i = e \mid F_i = f)$ is denoted by $\pi_{e|f(i)}$. For notational convenience we shall drop the subscript i when it is clear we are referring to an individual in the universe.

Let $XABC$ denote the cross-classification for the variables X, A, B, and C. For this table, define π_{xabc} as the expected proportion in cell (x,a,b,c) of the table. Using Bayes rule, we have the identity

$$\begin{aligned}\pi_{xabc} &= \pi_x \pi_{a|x} \pi_{b|xa} \pi_{c|xab} \\ &= P(X{=}x)\,P(A{=}a \mid X{=}x)P(B{=}b \mid A{=}a, X{=}x) \qquad (1) \\ &\quad \times P(C{=}c \mid A{=}a, B{=}b, X{=}x)\end{aligned}$$

Thus, the cell probability decomposes into a product of marginal and conditional probabilities. The true classification, X, is unobserved and will be treated in the subsequent analysis as a latent variable. Thus, the model is a type of *latent class model.*

The basic equations of the classical latent class model for two latent classes are

$$\pi_{abc} = \sum_{x=1}^{2} \pi_{xabc} \qquad (2)$$

where

$$\pi_{xabc} = \pi_x \, \pi_{a|x} \, \pi_{b|x} \, \pi_{c|x} \qquad (3)$$

(Goodman, 1974). This model postulates that the indicators A, B, C are mutually independent given the latent variable X. This assumption is referred to as "local independence."

Haberman (1979) demonstrated that (3) is equivalent to the hierarchical log-linear model

$$\log m_{xabc} = u + u_x^X + u_a^A + u_c^C + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC} \qquad (4)$$

where $m_{xabc} = N\pi_{xabc}$ for N = number of observations. The local independence assumption is implemented in the model by the exclusion of all interaction terms among the manifest variables A, B, and C. In what follows, the model in (4) will be denoted by {XA, XB,XC}. Since we will be dealing exclusively with hierarchical linear models, all terms of lower order than those in the braces are included in the model.

The local independence assumption is not likely to hold for the Census Test data since, as we have explained, the reconciliation process combines the results of the prior classification(s) with the current one. So, for example, if a person who is truly in the population ($X = 1$) is not listed by the Census ($A = 2$), they may be more likely to be classified as a nonresident by the ICM ($B = 2$); that is

$$P(B{=}2 \mid A{=}2, X{=}1) \geq P(B{=}2 \mid A{=}1, X{=}1) \qquad (5)$$

This would imply that the interaction (or correlation) between A and B given X is not zero. Likewise persons in the target population who are misclassified by the ICM may have a greater probability of being misclassified by the Evaluation Interview and, thus, the interaction term B, C, and X may be non-zero.

Therefore, we would be interested in fitting the models of the form

$$\begin{aligned}\log m_{xabc} = {} & u + u_x^X + u_a^A + u_b^B + u_c^C + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC} \\ & + u_{ab}^{AB} + u_{bc}^{BC} + u_{xab}^{XAB} + u_{xbc}^{XBC}\end{aligned} \qquad (6)$$

denoted by {XAB, XBC}. In this formulation, we still assume that the interactions XAC and between $XABC$ are 0. Other models are conceivable and will be explored in our analyses of the 1995 Census Test data.

Degrees of Freedom

The 2 x 3 x 3 table of manifest variables A, B, and C, contains 17 degrees of freedom and, thus, 17 parameters can be estimated from this table. The model (4) contains 11 parameters and can, therefore, be estimated leaving 6 degrees of freedom for testing the model lack of fit.

The model (6), however, contains 23 parameters and is, therefore, overspecified. Here we may use a device suggested by Hui and Walter (1980) for creating additional degrees by introducing groups. Let G denote an indicator variable for member in some group defined by one or more demographic variables. For example, let $G = 1$ for persons with Race = Black or Hispanic and $G = 2$ for Race = Other. Then, if the latent variable X and the manifest variable A, B, and C in (3) depend upon G, then (3) may be rewritten

$$\pi_{xgabc} = \pi_{gx} \, \pi_{a|gx} \, \pi_{b|gx} \, \pi_{c|gx} \qquad (7)$$

or, using log linear models with latent variables, as {GXA, GXB, GXC}. For two groups, the table $GABC$ provides 34 degrees of freedom and the model contains 22 parameters leaving 12 degrees of freedom for error.

The advantage for degrees of freedom of expanding the model to incorporate groups, however, can be seen by considering model (6). For example, if we assume that the manifest interaction terms are the same for both groups, then model (6) becomes {GXA, GXB, GXC, XAB, XBC}. It can be shown that this model contains 34 parameters and is thus estimable with the 34 degrees of freedom provided by table $GABC$. However, since the model is saturated, there are no degrees of freedom for testing the model fit.

5. Illustration of the Model for the 1995 Census Test Data

The $GABC$ table (i.e. cross-classification of Race, Census Classification, ICM Classification and Evaluation Interview classification) appears in Table 1. The first 2×3×3 table is the Census×ICM×Evaluation Interview table for a combined race category consisting of Blacks, Asian and Pacific Islanders (API), and Hispanics. The second 2×3×3 table is similar and combines all other races (including Whites). These data were weighted for the differential probabilities of selection across strata and then rescaled to the original sample size. Thus, the cell distributions reflect the population distributions in the combined study area while total sample size is maintained. This weighting will have some effect on the distribution the Pearson χ^2 statistics that we will use to determine model fit; however, our experience fitting these models to both the weighted and unweighted data gives us confidence that the model fit statistics for the weighted data are approximately χ^2 distributed for the data in Table 1.

It is important to note that what we refer to as the

Census classification is a misnomer since it does not represent the final classification of persons based upon the full Census mail enumeration and followup process. Because of the timing of the ICM, the full Census process had not been completed when the ICM was fielded and thus no Census roster information was available for some cases, particularly those in Stratum 5. Therefore, the misclassification rates we estimate for the so-called Census classifier are overstated and should not be interpreted as reflecting the accuracy of the Census enumeration and followup process.

Table 2 shows the results of fitting a number of models using the ℓEM Version 0.11 software (log-linear and event history analysis with missing data) developed by Jeroen Vermunt, Tilburg University, The Netherlands. Of the models considered, the best fit is provided by the model {GXA, XC, XAB} with 20 parameters and 16 degrees of freedom. The X^2 goodness of fit criterion is not rejected at the 5% level of significance (p = 0.087) indicating a reasonable fit of the model to the data.

The model postulates an interaction between Race and the Census classification, but no interactions between Race and the other two classifications. Further, the model postulates an interaction between the Census and ICM classifications; however, the interaction between the ICM and the Evaluation Interview classifications was not significant. Of course, all of the above significant effects depend upon the value to the latent variable X.

These test results indicate that the errors made by persons completing the Census form are related to Race. In Table 3 we provide for each group, the model estimators of the conditional probability of being classified in each category of the classifiers, A, B, and C given the true residency status. Note that for the Census classifier (A), the probability of being classified as "in the census" (A=1) for Blacks, API, and Hispanics in only 65.5% compared to 81.4% for Whites and others. For persons who are truly *not* in the population, Blacks, API, and Hispanics have a 21.7% chance of being classified as in the population (A = 2) versus a 10.1% chance for the Other race.

Now consider the ICM classifier (B). For persons who are truly in P the probability of being correctly classified is fairly high for both groups; viz, 88.9% and 92.3, respectively. However, the Black, API, and Hispanic group has a 7.8% chance of being missed (B=2) while the Other race has only a 4.8% chance. Both groups have a fairly high probability of being erroneously enumerated; i.e., classified as in P when they are truly *not* in P. The rate is 36.7% for Blacks, API, and Hispanics and 36.2% for the Other race. This would indicate that for nonresidents of the target population, the ICM classifier performs about the same for both race groups, which is generally poor.

Finally, for the Evaluation Interview classifier there is no difference by group. For persons who are in the population there is a near 0 probability of being classified as a nonresident (C=2), although 2.6% of this group may be classified as unresolved (C=3). Likewise, for persons who are truly *not* in P, there is a near 0 probability of being misclassified as a resident. A much larger proportion of this group, however, is classified as unresolved (53.5%). This can be a problem some of these unresolved cases are imputed as "residents" later in the undercoverage estimation stage.

6. Discussion

When the entire Census enumeration universe is considered, our analysis of the 1995 Census Test and Evaluation supports the Census Bureau's assumption that the Evaluation Interview is highly accurate: both false positive and false negative errors are approximately 0.

Our analysis of the 1995 Test Census and Evaluation demonstrates the usefulness of latent class models for the evaluating enumeration processes. Some of the benefits of this approach are that it:

- Allows testing of the assumptions associated with the measurement process; for example, the assumption of uncorrelated errors between measurements or infallible measurements can be formally tested.
- Provides estimates of the probabilities of false positive and false negative errors for all measures;
- Allows the testing of the equality of measurement effects across population domains and geographic strata; and
- Provides estimates of $P(X=1)$ for each cell of the cross-classification table $GABC$ (this feature was not demonstrated in the paper).

Thus, survey methodologists no longer have to rely on faith in the use of so-called gold standard measurements. The assumptions underlying the evaluation of measurement processes can be rigorously tested and maximum likelihood estimates of the error rates associated with the measures can be easily obtained using widely available software.

References

Goodman, L. (1974). "Analysis of Systems of Qualitative Variables When Some of the Variables are Unobservable. Part I: A Modified Latent Structure Approach," *American Journal of Sociology*, 79, 1179-1259.

Haberman, S. (1979). *Analysis of Qualitative Data, Vol. 2, New Developments*. Academic Press, NY.

Hui, S.L. and S.D. Walter (1980). "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 36, 167-171.

West, K. and Griffiths, R. (1996). "Results from the 1995 Integrated Coverage Measurement Evaluation Interview, paper presented at the 1996 Joint Statistical Meetings of the American Statistical Association, Chicago, Il.

Table 1. Cross-classification of Census ICM and Evaluation Interview by Race - Weighted Data

Note: "Black" denotes Blacks, Hispanics and API's; "White" denotes Whites and Others

Census Classification: A = 1 (Listed on Census Roster)							A = 2 (Not Listed on Census Roster)					
ICM Classification: B=1			B=2		B=3		B=1		B=2		B=3	
Race:	Black	White	Black	White	Black	White	Black	White	Black	White	Black	White
C=1	1119	722	12	13	34	15	463	128	123	36	29	7
C=2	5	1	5	2	2	1	20	3	22	21	5	2
C=3	42	14	5	0	7	1	40	11	36	13	4	2

Table 2. Test Results for Fitting Latent Class Models to G, A, B, and C

Model	X^2	*df*	$p(X^2)$	***n par***
XA XB XC	435.63	24	.000	12
GXA XB XC	48.76	20	.000	16
GXA GXB GXC	27.12	12	.007	24
GXA XB XC GXAB	14.01	8	.082	28
GXA XAB XBC	14.61	8	.067	28
GXA XC XAB	24.11	16	.087	20

Table 3. Estimated Classification Probabilities(Weighted Data) by Race-Group

	Blacks, Hispanics, and API		**Whites and Other**	
	Truly in P (X = 1)	**Truly NOT in P (X = 2)**	**Truly in P (X = 1)**	**Truly NOT in P (X = 2)**
Group proportions	**93.0**	**7.0**	**94.5**	**5.5**
Classifications	**Response Probabilities**			
A=1	65.5	21.7	81.4	10.1
A=2	34.5	78.3	18.6	89.9
B=1	88.9	36.7	92.3	36.2
B=2	7.8	51.3	4.8	54.0
B=3	3.3	12.0	2.9	9.8
C=1	97.4	0.0	97.4	0.0
C=2	0.0	46.5	0.0	46.5
C=3	2.6	53.5	2.6	53.5