MODELING CENSUS AND INTEGRATED COVERAGE MEASUREMENT PHASE MISSES IN THE CENSUS 2000 DRESS REHEARSAL

Michael Beaghen[1], Bureau of the Census
Washington, D.C., 20233

KEY WORDS: logistic regression, E-sample, P-sample.

I. Introduction

The Census 2000 Dress Rehearsal was conducted at three sites: Sacramento, California, Menominee, Wisconsin, and Columbia South Carolina and its surrounding counties. In this paper Columbia City is treated as a site separate from its surrounding counties, and Menominee is not included due to its small size. The methodology differed in the Columbia sites from the Sacramento site. In the Columbia sites a traditional census and a subsequent Post Enumeration Survey (PES) were performed. In Sacramento the Integrated Coverage Measurement (ICM) was performed which consisted of two phases: an initial phase akin to a traditional census and a secondary survey akin to a PES. For the purposes of this analysis the methodologies are treated the same. The primary survey shall be referred to as the census and the secondary survey shall be referred to as the ICM. The census enumerated the entire sites. After the census was completed, the ICM performed an independent reenumeration conducted only in selected block clusters within the sites. In Sacramento a single estimate was produced based on the census and the ICM. In South Carolina an estimate based on the census and an estimate based on the PES was produced.

For the ICM sample the block clusters were the primary sampling units. They were drawn from twelve sampling strata. Those selected block clusters with 80 or more housing units were subsampled. The selected block clusters after subsampling comprised the ICM sample.

The census enumerations within the ICM sample block clusters define the E-sample. The people enumerated in the ICM in the sample block clusters define the P-sample. A matching operation linked E-sample people with P-sample people. A linked pair is called a match. An E-sample enumeration that was not linked to a P-sample person was an E-sample non-match. A follow-up interview determined whether the person existed in the ICM sample. People found to exist are called confirmed non-matches. People found not to exist in the sample are called erroneous enumerations. Erroneous enumerations will be ignored in this analysis. Confirmed E-sample non-matches represent P-sample misses or failures to capture.

A P-sample person that does not match to an E-sample enumeration is called a P-sample non-match. A follow-up interview determined whether the person existed in the ICM sample. Confirmed P-sample non-matches represent E-sample misses or failures to capture.

The purpose of this paper is to use logistic regression models to relate these P-sample misses and E-sample misses to demographic characteristics and housing unit characteristics. The limitation of univariate descriptive statistics is that they do not address the question of the relationship of one variable in the context of other variables. A regression type model avoids this limitation. Since the response is binary, that is, a person is either captured or missed, logistic regression is an obvious method.

This study is observational rather than experimental. The characteristics used as regressors in the model are not controlled by the researcher but rather are random variables. Consequently the modeling is not predictive but descriptive and the hypothesis tests used to determine which variables to include in the model are not strictly correct. They are to be understood as guidelines in model building.

The paper is organized as follows. First, the variables are laid out and described. Then I build two models. To model the P-sample misses, I model the E-sample people who were matched to ICM people against those who were confirmed non-matches. The erroneous

[1]Michael Beaghen is a mathematical statistician in the Decennial Statistical Studies Division of the US Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. The research results and conclusion are expressed are those of the author and do not necessarily indicate concurrence by the Census Bureau. It is released to inform interested parties of current research and to encourage discussion.

enumerations were discarded from the analysis. To model the E-sample misses, I model the P-sample people who were matched against those who were confirmed non-matches.

II. The Variables

In both analyses the response variable has two outcomes: a person is a match or a person is a non-match that is confirmed to exist.

The values of the regressor variables in both data sets were formatted such that they would have the same categories, allowing the models to be comparable. All of the regressor variables are dichotomized. Categorical variables indicating group membership (in italics) are represented as dichotomous indicator variables, with one category serving as a baseline to which all the other categories are compared. For example, for the variable race, white serves as the baseline. Age, the only continuous variable, was formatted as a categorical variable. The breakdown of the categorical variables into classes is as follows.

Race: *black* (black or African-American), *Asian* (Asian or Pacific Islander), *mixed* (mixed ancestry, American Indian or other); white is the reference group.

Age: *child* (age <= 17), *young* (18 <=age <= 29), *middle* (30<=age<=49); older (50 <=age) is the reference group.

Tenure: *renter* (all nonowners); owner is the reference group.

Site: *Columbia, RuralSC*; Sacramento is the reference group.

Relationship to Respondent:
family (person was an immediate family member of respondent), *relative* (person was an non-immediate family member of the respondent), other (a non-relative, or relative beyond immediate family or cousins, grandparents, or aunts or uncles); the respondent serves as the reference group.

Multi-unit:
multi-unit (person lived in a multi-unit); the reference group is formed by those living in single units, which includes trailers.

III. Methods

I used SUDAAN's (Shah, Barnwell & Bieler, 1997) Proc Logistic to estimate the models. SUDAAN is designed to handle data from complex surveys such as the ICM.

I also used SAS (1989) software's Proc Logistic to estimate the models. SAS has more modeling capabilities than SUDAAN. To account for the complex survey design two measures were taken. First, the observations were weighted by the inverse of the sampling weight. This made the estimates reflect the population of the sites, not the sample. Second, the Wald test statistic was divided by a design effect since SAS estimates variances as if the sample were a simple random sample. This method gives the correct maximum likelihood estimates but the Wald test statistics are approximate at best.

The SUDAAN modeling is viewed as more correct and generated the estimates seen in the tables.

The backward, forward and stepwise methods of model selection yielded similar models, though the backward selection models were superior for both models. The cutoff for removing a variable was a p-value greater than 0.03. This choice was somewhat arbitrary. It allowed me not to include some variables that had only very weak effects.

IV. Interpretation - Modeling the P-Sample Misses

In logistic regression, the binary response is thought of as successes versus failures. In this model define a success as a person in the E-sample but not on the P-sample, i.e., a P-sample miss. Define a failure as a match. Rather than examining the parameter weights themselves it is easier to interpret the odds ratios associated with an increase of one unit for each parameter. These odds ratios are directly related to the parameter weights. Since each of the variables has a value of zero or one, depending on group membership as described in Section 2., the interpretation of the odds ratios are straightforward. The odds ratio will refer to the ratio of the odds of a response with a value of one with the odds of a response with a value of zero. For example, see the variable *renter* in Table 1. *renter* has an odds ratio of 1.48. Since a value of one indicates a person lives in a rented unit and a value of zero indicates a person lives in an owned unit, the odds ratio shows that a *renter* has a 48% greater odds of being missed by the P-sample than an owner, all other variables being held constant.

If the variable of interest is an indicator variable in a

series of indicator variables that denote membership to a group the odds ratio refers to the comparison with the baseline group (Hosmer, Lemeshow 1989). As an illustration, consider race. As shown in Section 2., the four race categories, white, black, Asian, and mixed, are indicated by three indicator variables, *black*, *Asian*, and *mixed*. Only *black* and *mixed* are in the model. The odds ratio of 1.31 associated with those of mixed ancestry implies that a person of mixed ancestry has a 31% greater odds of being found in the E-sample and missed in the P-sample than those a person of white ancestry.

Another factor to consider in assessing the relative importance of a variable is how often it has a value of one. For example, although *black* has an odds ratio a good bit higher than *male*, there are only a fraction as many people who are *black* as *male*.

To summarize the results of the model, race, tenure, multi-unit, age, relationship to respondent, sex and blocksize all are associated with capture in the P-sample. Renters are more likely to be missed than are non-renters. Non-immediate relatives and non-relatives are more likely to be missed than the respondent or the respondent's immediate family (i.e., the variable indicating immediate family was not significant). People of black or mixed ancestry are more likely to be missed than those of white or Asian ancestry. People who live in multi-units are more likely to be missed than people who live in single units. Children and young adults are more likely to be missed than middle aged or older people. Males were slightly more likely to be missed than females.

Also illuminating are the several categorical variables that are non-significant. They are Hispanic origin and all second level interaction terms.

Table 1. Odds Ratios

Reference Variable	**Model Variable**	**P-Sample Misses Odds Ratios**	**E-Sample Misses Odds Ratios**
Owner	*Renter*	1.48	1.32
Non-Hispanic	*Hispanic*	ns	ns
Female	*Male*	1.09	1.05
Older	*Child*	1.58	1.21
	Young	1.24	1.33
	Middle Age	ns	ns
Respondent	*Relative*	1.87	1.36
	Family	ns	ns
	Other	1.56	1.68
White	*Black*	1.31	1.54
	Asian	ns	1.43
	Mixed	1.31	1.36
Single-Unit	*Multi-Unit*	1.59	1.73
Sacramento	*RuralSC*	ns	1.56
	Columbia	ns	0.65
Non-Large	*Large Block*	1.39	ns

V. Modeling the E-Sample Misses

A P-sample confirmed non-match represents a person missed by the census, that is an E-sample miss. In this second model define a success as an E-sample miss and define a failure as a match. I will compare this model, which models census misses, to the model of the previous section which modeled the P-sample misses. It is reasonable to ask if the people missed in the ICM and the Census are similar in their characteristics because both eluded similar surveys. Upon examination of the odds ratios this seems to be true to an extent, though there were some important differences in the models.

Firstly, a person in the Columbia city site was less likely to be missed by the census than a person in the Sacramento site, though a person in the rural South Carolina site had the greatest chance of all of being missed by the census. In comparison, the P-sample site was not related to the miss rate.

Secondly, P-sample people were more likely to be missed in large blocks, though this was not true for E-sample people.

Except for the three situations just described, the characteristics describing E-sample misses are similar to those describing P-sample misses. Age, relationship to respondent, race, tenure and multi-unit are similar in the nature of their association with misses. Likewise, in both surveys sex and Hispanic origin played little or no role. See Table 1. which shows the odds ratios for each variable.

VI. Conclusion

Logistic regression is a useful method to examine what variables are associated with P-sample and E-sample misses. Not surprisingly, many of the same variables associated with P-sample misses are associated with E-sample misses. They are age, race, tenure, multi-unit and relationship to the respondent. The role site played differed in the ICM and census. Hispanic origin played no role and sex played a small but similar role in either census or ICM capture. Second order interactions were not statistically significant. This suggests that had I modeled separately for Sacramento and South Carolina that I would have similar results.

This work could have implications for the poststratification. Poststrata are defined such that the probability of capture is as homogeneous as possible within each poststratum, for both P-sample capture and E-sample capture (Wolter 1986). In the 1998 ICM poststratification was done by tenure, age, sex, race and Hispanic origin. The results here suggest different poststrata. Sex and Hispanic origin were of little use in discriminating capture probabilities, relationship to respondent and multi-unit status did better in discriminating capture probabilities.

Also, the comparison between using SUDAAN and SAS is of interest. The parameter estimates were the same for practical purposes. However, the Wald statistics and the p-values differed. In the SAS models the variables *Male* and *Asian* were not statistically significant. The design effect used for SAS was constant for all variables, though the SUDAAN output clearly shows that the design effects vary for the variables.

VII. References

Childers, Danny R. (1998): *The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement.* DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter F-DT-2

Hosmer, David W., Lemeshow, Stanley (1989): *Applied Logistic Regression.* John Wiley & Sons, New York.

Waite, Preston Jay, and Hogan, Howard (1998): *Statistical Methodologies for Census 2000 Decisions, Issues and Preliminary Results.* Presented at the Joint Statistical Meetings, Session on Social Statistics, August 13, 1998, in Dallas, Texas.

Wolter, Kirk M. (1986): *Some Coverage Error Models for Census Data.* Journal of the American Statistical Association, June 1986, Vol. 81, No. 394.

SAS (1994): *SAT/STAT User's Guide, Version 6, Fourth Edition Volume 2.* The SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513

Shah, B.V., Barnwell, B.G., and Bieler, G.S. (1997): *SUDAAN, User's Manual Volume II, Release 7.5.* Research Triangle Institute, Research Triangle Park, NC 27709