2010 Census Coverage Measurement - Initial Results of Net Error Empirical Research using Logistic Regression

Richard Griffin, Thomas Mule, Douglas Olson[1]
U.S. Census Bureau

1. Introduction

This paper reports the initial research results for 2010 Census Coverage Measurement (CCM) of net error estimation using logistic regression models. For the dual system estimates of coverage error in past censuses, a post-stratification approach has been used. The post-stratification approach has some significant limitations since it limits the number of factors that can be included because each factor added can crudely be thought of as cutting the post-stratum sample sizes in half. Statistical modeling techniques like logistic regression potentially offer more flexibility and possibilities for reducing sampling error, synthetic error, and correlation bias in the estimation.

The initial work used a limited set of variables which will be expanded as the research evolves. The first phase had three goals:

1. Gain experience using SAS software to implement necessary computations for regressions and population estimators.

2. Investigate the trade off between bias and variance of estimates obtained by the elimination of higher order interaction terms in the models

3. Examine measures to evaluate and compare the fit of alternative models.

Section 2 discusses background references and Section 3 describes the data used. Section 4 describes the variables included in each of the models examined in this paper. Section 5 gives detailed methodology; sub-sections include logistic regression (5.1), model selection measures (5.2), population estimation alternatives (5.3), and standard errors (5.4). Section 6 provides results and section 7 provides a summary and future work. Section 8 details the references.

2. Background

Griffin (2005) lays out an approach for using logistic regression modeling instead of post-stratification for the estimation of net errors. The basis for the logistic regression approach is the final report on model-based estimation of population size prepared by the National Opinion Research Corporation (NORC) for the U.S. Census Bureau (Habermann et. al (1998)). Their research used the 1990 Post Enumeration Survey (PES) data (Hogan (1992, 1993). Habermann et al. used separate logistic regressions of the correct enumeration status of the E sample and the match status of the P sample. Building on the logistic regression results, they suggested five possible estimators as Population Estimation Alternatives.

3. Data

Data collected for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) is used. The E-sample consists of census data defined persons in A.C.E. sample blocks. The P-sample consists of independently enumerated persons in these same sample blocks. See U.S. Census Bureau (2003).

In order to correct for the measurement errors detected in the original March 2001 results, the original estimation methodology was adapted. The new methodology allowed the estimate of correct enumerations from the E sample and the estimates of matches and P-sample totals from the P sample to be adjusted. When creating the source files for potential research, we wanted to come up with a way to allocate these aggregate corrections to the individual E and P-sample cases. We also wanted this allocation to be done in a way that using information on these files would produce approximately the same results as the A.C.E. Revision II estimates. For simplicity, we were also interested in an allocation so that we could use the original March 2001 formulas that would produce similar results to the A.C.E. Revision

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

II. We decided to do this by creating new versions of the sampling weights and correct enumeration, match and residence results. Using these variables allowed us to do our research using the full E and P samples while still accounting for the most of the adjustments employed in the A.C.E. Revision II work.

There were a few differences between the A.C.E. Rev. II calculations and those employed in the current work:

- No adjustments were made for correlation bias.
- Only nonmover and outmover cases were used in the calculation of match rates and the PES-A formula was used for determining the weighted match and P-sample total quantities.
- No "possible conversion to mover" adjustment was used.

Almost all cases of correct enumerations and matched persons in the Research File have had their Correct Enumeration (CE) or Match probability adjusted to slightly less than one (e.g. a Matched person whose match probability had been 1.0000 is now .985 of a Match). To perform logistic regressions, most persons have been classified into a CE or Match part and an Erroneous Enumeration (EE) or Nonmatch part by proportional-izing their weight. (For instance, a person who is 99.1 percent of a CE and whose weight was 100 is now 99.1 weighted CE's and 0.9 weighted EE's.) Mathematically, this does not affect point estimates of Correct Enumeration or Match rates of population groups and has only a trivial effect on variances.

Mule and Olson (2005A) provides more information about the appending of the A.C.E. Revision II coding and missing data variables from the revision files for each person onto the full person files for the E and P samples.

4. Models

We started our analysis by deciding to use models that included only the variables used in the March 2001 post-stratification. See Griffin (2000) for more information on the post-stratification. The following variables were used as part of the post-stratification:

- Race/Origin domains (7 groups)
- Age/Sex groupings (7 groups)
- Tenure (Owner, Non-Owner)
- MSA/TEA classifications (4 groups)
- Region (4 groups: Only for Non-Hispanic White and Other Domain Owners)
- Mail Return Rate (High or Low areas: Different areas for Non-Hispanic White and Other, Non-Hispanic Black and Hispanic Domains)

In this initial research, we examined six models that used combinations of these variables. These variables were used to run separate logistic regressions of the correct enumeration and match status. Models are identified by the number of parameters.

1. **416** Collapsed Post-strata

The first model used the same 416 collapsed post-strata that were used for the March 2001 estimation. This will serve as a baseline showing an example of the post-stratification methods used in the past. This can also be considered as 415 individual dummy variables in a logit model.

2. March 2001 First Order Interactions **(150 parameters)**; and

3. March 2001 Main Effects **(23 parameters)**

Details of how the March 2001 variables were used in the main effects model as well as the first order interaction model are given in Mule and Olson (2005B).

4. ROAST **98**

Our next model used only the Race/Origin domains, Age/Sex groupings and Tenure. This is a fully saturated model using all 98 cross-classifications of these 3 variables. We will use the term "ROAST" to refer to models using these 3 variables. This is another example of post-stratification but with fewer variables than Model 1. Similar to Model 1, this could be considered as 97 individual dummy variables in a logit model.

5. ROAST First Order Interactions **(62 parameters)**

Our next model uses the ROAST variables with the main effects and the first order interactions. Including the intercept, there are 62 parameters in this logistic regression model.

6. ROAST Main Effects **(14 parameters)**

Our next model uses the ROAST variables but only as main effects in the logistic regressions. Including the intercept, there are 14 parameters in this logistic regression model.

5. Methodology

5.1 Logistic Regression

We modified SAS Interactive Matrix Language (IML) code to do separate logistic regressions for correct enumeration and match status for each of the 6 models. The following describes the weighted logistic regression in general and how we accounted for probabilities of correct enumeration and match status between 0 and 1. For the two regressions, a "correct enumeration" or a "match", respectively, are considered a successful outcome. These logistic regressions used the adjusted sampling weights and probabilities from the A.C.E. Research File so results similar to A.C.E. Revision II without the correlation bias adjustment could be obtained by using just the full E and P samples.

The dependent response variable is 1 for a success and 0 for a failure. Two records were created for each person. One record is given a dependent response value of success and a weight equal to the product of the adjusted sampling weight and the adjusted probability of success (correct enumeration or match). The second record is given a dependent response value of failure and a weight equal to the product of the adjusted sampling weight and the adjusted probability of failure (erroneous enumeration or nonmatch). The adjusted probability of failure is equal to 1 minus the adjusted probability of success.

E-sample persons with insufficient information for matching were included as erroneous enumeration cases in the modeling. This is different than Haberman et al. as they removed these cases from the E-sample (i.e., treated them the same as whole person imputations).

In the research contained in this paper, population groups were created from the post-stratification used on the March 2001 A.C.E. estimates, because population totals from the Census were readily available. Additional research will employ additional population group totals, which will require totaling the Census for all groups created.

5.2 Model Selection Measures

This section describes the measures used in our initial research to evaluate and compare the performance of the logistic regressions of the models listed above. In our initial research, logarithmic penalty functions, jackknife estimates of bias of this function and cross-validation were used.

Logarithmic Penalty Function

In order to assess the performance of each of the 6 models in the logistic regression analysis, we started with the logarithmic penalty function that was used by Habermann et al. in their previous research. They used this measure to assess the predictive ability of each of the models.

The logarithmic penalty function for the correct enumeration status is

$$\hat{H}_E = -\frac{1}{W_E}\Big[\sum_{j \in E\,sample} w_{e(j)} p_{ce(j)} \ln\left(\hat{\pi}_{ce(j)}\right) + w_{e(j)}\left(1 - p_{ce(j)}\right) \ln\left(1 - \hat{\pi}_{ce(j)}\right)\Big]$$

Where W_E is the weighted total for the E sample, $w_{e(j)}$ is the adjusted sampling weight, $p_{ce(j)}$ is the adjusted correct enumeration probability, and $\pi_{ce(j)}$ is the predicted correct enumeration probability from the model.

The logarithmic penalty function for the match status is

$$\hat{H}_P = -\frac{1}{W_P}\Big[\sum_{j \in P\,sample} w_{p(j)} p_{m(j)} \ln\left(\hat{\pi}_{m(j)}\right) + w_{p(j)}\left(1 - p_{m(j)}\right) \ln\left(1 - \hat{\pi}_{m(j)}\right)\Big]$$

Where W_p is the weighted total for the P sample (nonmovers and outmovers), w_p is the adjusted sampling weight, $p_{m(j)}$ is the adjusted match probability, and $\pi_{m(j)}$ is the predicted match probability from the model.

Jackknife Estimate of Bias of the Logarithmic Penalty Function

Habermann et al. bring up the issue of adjusting their log penalty measure because of the bias in the statistic. They used the following jackknife approach to estimate the bias.

The jackknife bias estimator is:

$$Bias_{Jack} = (g-1) \times \left(\hat{\theta}_{(.)} - \hat{\theta}\right)$$

Where $\hat{\theta}_{(.)} = \frac{\sum_{i=1}^{g} \hat{\theta}_i}{g}$,

$\hat{\theta}$ is the full sample estimate of the log penalty function

$\hat{\theta}_i$ is the *i*th replicate estimate of the log penalty function and g is the number of groups(replicates).

Cross-Validation

The National Academy of Science Panel on Coverage Evaluation and Correlation Bias in the 2010 Census suggested using cross-validation as a model assessment tool. Chauchat et al (2002) suggest a cross-validation approach for clustered data. Our research used a *k*-fold cross-validation methodology where our sampled clusters were sorted by cluster number and systematically assigned to *k* groups.

A k-fold cross-validation of a model is implemented by the following steps:

1. The sample data were randomly assigned into *k* groups

2. The logistic regression of the correct enumeration rate or the match rate were applied to the entire sample except for one *k* part. The estimated logistic regression parameters were obtained.

3. Using a) the parameters estimated in Step 2 and b) the sample in the *k*th part., the log penalty function (LP) was estimated.

4. This was repeated for each of the *k* groups.

5. A generalized rate was estimated by

$$LP_{cross-validation} = \frac{\sum_{i=1}^{k} LP_i}{k}$$

This generalized rate is biased but the bias becomes negligible when *k* becomes large. The random variation of the generalized rate increases and the calculation time increases with *k*. The random variation increases because as *k* increases, each group then has fewer cases contributing to the group estimate and thus the variability increases. Our research explored various numbers k of groupings to check the sensitivity of the choice.

5.3 Population Estimation Alternatives

The next step is to use the models and the post-stratification or regression results to be able to generate estimates of the population. Habermann et al. (1998) suggested five estimators to do this.

All of the estimators are functions of the following three quantities:

- Data-defined enumerations in the Census (not including reinstated records)
- Correct enumeration probability
- Match probability

This section gives the formula of each estimator and the data and information used in each. The formula explanations are for national estimates. Results for subpopulations can be obtained with little modification by summing only over cases for that subpopulation.

N1 Estimator

The N1 estimator uses all of the data-defined enumerations in the census (not including reinstated cases). Based on results of the modeling and the characteristics of each case, we can estimate a predicted probability of the correct enumeration and match status.

The formula for the N1 estimator is:

$$\hat{N}_1 = \sum_{j \in C_{DD}} \frac{\hat{\pi}_{ce(j)}}{\hat{\pi}_{m(j)}}$$

where C_{DD} is the data-defined enumerations in the Census (not including reinstates), $\pi_{ce(j)}$ is the predicted correct enumeration probability from the model and $\pi_{m(j)}$ is the predicted match probability from the model.

N2 Estimator

The N2 estimator uses only the sample data. The data-defined records in the census (not including reinstated cases) are accounted for by the E sample. Based on results of the modeling and the characteristics of each E-sample case, we can estimate a predicted probability of the correct enumeration and match status. This estimator may be more appealing than the N1 estimator if good covariates are only available for the sampled cases and not for all of the enumerations in the census. This may be more beneficial in future research when additional variables are explored.

The formula for the N2 estimator is:

$$\hat{N}_2 = \sum_{j \in ESample} w_{e(j)} \frac{\hat{\pi}_{ce(j)}}{\hat{\pi}_{m(j)}}$$

where $w_{e(j)}$ is the adjusted sampling weight of the E-sample case, $\pi_{ce(j)}$ is the predicted correct enumeration probability from the model and $\pi_{m(j)}$ is the predicted match probability from the model.

N3 Estimator

The N3 estimator is similar to the N2 estimator since it too only uses the sample data. The probability of correct enumeration of each E-sample case is used instead of the predicted value from the modeling. A predicted probability of the match status is estimated for each E-sample person. If using only sample data, this estimator may be more appealing than the N2 estimator since erroneous enumerations in the sample will be assigned a zero probability of correct enumeration.

The formula for the N3 estimator is:

$$\hat{N}_3 = \sum_{j \in ESample} w_{e(j)} \frac{\hat{p}_{ce(j)}}{\hat{\pi}_{m(j)}}$$

where $w_{e(j)}$ is the adjusted sampling weight of the E-sample case, $p_{ce(j)}$ is the correct enumeration probability of the E-sample case and $\pi_{m(j)}$ is the predicted match probability from the model.

N2R Estimator

The N2R estimator is the N2 estimator where the weighted estimates of the data-defined enumerations from the E sample is ratio adjusted to a census count of data-defined persons. This helps reduce the bias and variance of the population estimates.

N3R Estimator

The N3R estimator is the N3 estimator where the weighted estimates of the data-defined enumerations from the E sample is ratio adjusted to a census count of data-defined persons. This helps reduce the bias and variance of the population estimates.

5.4 Standard Errors

Standard errors of all estimates were computed using a jackknife methodology that used 100 groupings. The 100 random groupings were assigned using the last two digits of the A.C.E. cluster number including the check digit.

6. Results

Detailed results are given in Mule and Olson (2005B). This section provides a summary of the results.

Model Selection Measurements

Table 1, at the end of this paper, shows the logarithmic penalty function, jackknife bias and cross-validation measures for the 6 models. Results are shown for both the correct enumeration and match regressions.

As expected, the logarithmic penalty function results show that the penalty function decreased as the number of parameters increased. All differences were statistically significant at the .001 (0.1%) level. Haberman et al. suggested that differences in the logarithmic penalty functions of 0.01 are substantial

and differences of 0.001 are rather small. Research is ongoing to evaluate this suggestion. However, using this suggestion, although all differences are statistically significant, the differences are rather small. Note also that a bias correction applied to correct for overfitting would make these differences even less meaningful.

For both regressions, we are seeing different ordering using the cross-validation measures as compared to the ordering of the log penalty measure from the full sample. For correct enumerations, the cross-validation measure of Model 2: March 2001 First Order Interactions is showing a lower estimate as compared to the Model 1: 416 post-strata estimate. For matches, the cross-validation of both Model 2: March 2001 First Order Interactions and Model 3: March 2001 Main Effects have a lower estimate as compared to the Model 1: 416 post-strata estimate. The ordering of the logarithmic penalty functions for both the correct enumeration and match rate do not change when the penalty estimates are adjusted for the jackknife bias estimate.

Population Estimates

Tables 3 - 7 of Mule and Olson (2005B) show detailed population estimates and their standard errors. Due to page restrictions only a summary is provided in this paper.

For most domain/tenure combinations the standard errors of the estimates increased as more parameters were added to the model. The opposite relationship was seen for American Indian on Reservation Non-owners and Hawaiian and Pacific Islander Non-owners. The standard errors of Hawaiian and Pacific Islander Owners are lower for the 416 post-stratification than for the ROAST 98. This happened because the 416 production model collapsed the 7 age/sex categories for this domain/tenure combination into 3 categories while it remained 7 for the ROAST 98. For Non-Hispanic Black non-owners the standard errors remained relatively constant even though more parameters were added.

The coverage correction factor (CCF) point estimates are impacted by the different models. One example is American Indian on Reservations. The addition of more variables and parameters in the models increases the CCF for owners but decreases the CCF for non-owners. The standard error for national totals decreased as more parameters were added to the model, opposing the trend observed in most Domain/Tenure groups. More research is needed on this seeming contradiction.

The N2R and N3R estimators, that ratio adjust the results using the E-sample data to the data-defined counts, produces point estimates and standard errors similar to those for the N1 estimate. The CCF estimates using the N2 and N3 estimator, especially for the American Indians on and off reservation estimates, are very different as compared to the N1 estimate. As expected the standard errors for the N2 estimates are much larger than those for the N2R estimates and the standard errors for the N3 estimates are much larger than those for the N3R estimates. We would not use either N2 or N3 since N2R and N3R, which use a ratio adjustment, are better in terms of bias as well as variance.

7. Summary and future work

This work has given us confidence that we can implement estimation of net error using logistic regression modeling techniques and that these models have the potential to improve net error estimation. Future work will look at using other variables including those identified for the E-sample post-stratification in Accuracy and Coverage Evaluation (A.C.E.) Revision II.

8. References

Chauchat, J.H., Rakotomala, R. And Pelligrino, F. (2002) "Error Rate Estimation for Clustered Data - An Application to Automatic Spoken Language Identification," Proceedings of Statistics Canada Symosium.

Griffin, Richard (2000) "Accuracy and Coverage Evaluation Survey: Final Post-stratification Plan for Dual System Estimation," DSSD Census 2000 Procedures and Operations Memorandum Series Q-24, U.S. Census Bureau, April 19, 2000.

Griffin, Richard (2005) "Net Error Estimation for the 2010 Census," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-01, U.S. Census Bureau, April 18, 2005.

Habermann S.J., Jiang, W. And Spencer B.D. (1998), "Activity 7: Develop Methodology for Evaluating Model-Based Estimates of the Population Size for States Final Report," prepared by NORC for the U.S. Census Bureau under contract no. 50-YABC-2-66023.

Hogan, H. (1993) "The 1990 Post-Enumeration Survey: Operations and Results", *Journal of the American Statistical Association*, 88, 1047-1060.

Hogan, H. (1992) "The 1990 Post-Enumeration Survey: An Overview", *The American Statistician*, American Statistical Association, Alexandria, VA. 261-269.

Mule, Thomas and Olson, Douglass (2005A) "A.C.E. Revision II - Computer Specifications for Research Files of A.C.E. Revision II Person Data," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-02, U.S. Census Bureau, April 18, 2005.

Mule, Thomas and Olson, Douglas (2005B) "A.C.E. Revision II - Initial Results of Net Error Empirical Research using Logistic Regression," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-03, U.S. Census Bureau, April 18, 2005.

U.S. Census Bureau (2003b) "Technical Assessment of A.C.E. Revision II" March 12,2003. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/ACETechAsse ss.pdf

Table 1: Model Assessment Results

	Model	Parameters (including intercept)	Correct Enumeration			Match		
			Log Penalty Estimate	Jackknife Bias Estimate	Cross-Validation	Log Penalty Estimate	Jackknife Bias Estimate	Cross-Validation
1	416	416	0.23396	-0.00043	0.23480	0.25809	-0.00086	0.25987
2	March 2001 First Order Interactions	150	0.23427	-0.00020	0.23462	0.25857	-0.00041	0.25935
3	March 2001 Main Effects	23	0.23507	-0.00005	0.23511	0.25928	-0.00009	0.25938
4	ROAST 98	98	0.23511	-0.00008	0.23521	0.26102	-0.00015	0.26126
5	ROAST First Order Interactions	62	0.23515	-0.00006	0.23521	0.26112	-0.00012	0.26130
6	ROAST Main Effects	14	0.23569	-0.00003	0.23568	0.26145	-0.00005	0.26148

Note: 20 grouping results shown for jackknife bias and cross-validation measurements