A MODEL FOR IMPUTING NONSAMPLE HOUSEHOLDS WITH SAMPLED NONRESPONSE FOLLOW-UP

Elaine Zanutto, Harvard University and Alan M. Zaslavsky, Harvard University
Elaine Zanutto, Department of Statistics, 1 Oxford Street, Cambridge MA 02138

KEY WORDS: Census, Loglinear Models, Iterative Proportional Fitting

1 Introduction

Sampling for nonresponse follow-up (NRFU) is a probable innovation for census methodology in year 2000. The potential cost savings for NRFU sampling are large, but it is necessary to show that we can attain an acceptable level of accuracy for small areas before such a sampling scheme can be adopted.

The following is a brief description of data collection under NRFU sampling. At the first stage, census data are collected by mailout-mailback (possibly in combination with other methodologies such as a truncated field/telephone follow-up operation) in an area (say, a District Office (DO)). At the second stage, follow-up (field or telephone) is carried out for a sample of the nonresponse cases from the first stage. The sample consists of all nonresponding addresses in a sample of the blocks in the area. (Reasons for requiring the NRFU sample to be a sample of blocks rather than individual households involve the interaction of NRFU sampling with coverage measurement and the exigencies of field management.) Second stage follow-up is assumed to be complete in the sample blocks, meaning that all addresses either are resolved to be vacant or are resolved by completing a questionnaire for the household that lives there.

The problem is to estimate/impute the characteristics of households at addresses in nonsample blocks from which no response was obtained at the first stage. Once the census roster is completed by imputation, all tabulations prepared from the completed roster are guaranteed to be consistent with each other.

Current work on this problem follows one of two basic strategies. Isaki, Tsay, and Fuller (1994) and the authors of this paper pursue what might be called a "top-down" strategy which starts with aggregates of households and subdivides them in a manner that maintains consistency with estimates calculated at the aggregate level. Simple ratio models (as in Isaki, Tsay, and Fuller) or more complex loglinear models (as in this paper) are used to estimate counts for small areas and detailed demographic groups for which direct estimates are not possible. These ad hoc models do not describe the full complexity of the units but they are designed to maintain consistency of the aggregates which are considered most important.

Schafer (1995) develops a "bottom-up" strategy in which households are built up from individual persons and their characteristics and relationships, each of which must be described by its own model. This strategy gives a more complete and detailed description of the population and, if carried out successfully, it can support full probability (e.g. Bayesian) inferences about its unobserved characteristics. However, this approach, unlike the other, requires that a fairly complex set of models be built before any imputations can be made. Furthermore, in this framework it is more problematic to maintain consistency between microdata and aggregate controls. A combined strategy, however, could use our models to produce nearly unbiased estimates by types and Schafer's models to complete the imputations.

The general framework of the sampling and estimation procedure assumed in this paper is as follows:

1. Blocks are sampled according to the selected sampling scheme and rate.
2. A model is fit using the respondents and the sampled nonrespondents.
3. Predicted counts are calculated for each block.
4. Counts are rounded.
5. Households are imputed for each block.
6. The completed rosters are used to prepare tabulations and microdata samples.

This general procedure is similar to that described in Isaki, Tsay and Fuller. The difference is the form of the model that is used in steps 2 and 3. Isaki, Tsay and Fuller use a stratified ratio model whereas we use a loglinear model.

In this paper, we focus on steps 2 and 3 of this process in order to explore, through simulations, the gains in accuracy that are possible with increasingly sophisticated models. This work builds on earlier work by Zanutto and Zaslavsky (1994, 1995).

2 The model and its interpretation

We use the following notation for geography and demographic characteristics of households: i = block index, $a = a(i)$ = Address Register Area (ARA) for the ARA containing block i, where the "ARA" is a generic term for an area intermediate in size between a block and the entire area under consideration (DO), j = index of household type, $x_1 = x_1(j)$ = set

of covariate values associated with household type j, x_2, x_3, x_4 = other sets of covariate values associated with household type j, where x_2 and x_3 are each assumed to be coarser than (expressible as linear functions of) x_1, and x_4 is assumed to be coarser than x_2 and x_3, and r = first-stage (mail) response indicator where $r = 0$ for responding households and $r = 1$ for nonrespondents.

We assume a loglinear model of the form:

$$\log \mathbf{E}n(i,j,r) \sim i+x_1+r+i*x_2+i*r+r*x_3+r*a*x_4$$

where the left hand side is the logarithm of the expected count for a given block, household type and response status, and the right hand side represents a linear predictor determined by the covariate values, the response indicator r, and the indices i and $a = a(i)$.

This model is motivated by the following principle of maximum likelihood estimation in loglinear models: In a hierarchical loglinear model (i.e. one in which for every interaction effect, all main effects or interactions marginal to it are also included in the model), the expected values for every margin corresponding to an effect in the model are equal to the corresponding observed margins. Therefore, since each of the terms in this model can each be interpreted as a margin of the block×type×response table, if we fit the model by maximum likelihood, the expected (fitted, predicted) values for these margins will match those observed in the data.

More detailed interpretation of the terms of the model, motivation for the model, and discussion about the covariates $x_1, \ldots, x_4$ and ARA indices a are given in Zanutto and Zaslavsky (1994, 1995).

3 Fitting the model and calculating imputations

We fit the model using the iterative proportional fitting (IPF) algorithm. The IPF algorithm successively adjusts fitted cell counts so they match each marginal table in the set of minimal sufficient statistics for the model. This iterative procedure continues until the maximum difference between the sufficient statistics and their fitted values is sufficiently close to zero. These estimates converge to the maximum likelihood estimates. The unobserved cells (i.e. we have information on responding households in all blocks but for nonresponding households only in sample blocks) are not a problem because nonsample nonrespondents contribute to the likelihood only through the total number of nonrespondents in each block. Therefore, to maximize this part of the likelihood we need only ensure that the fitted number of nonrespondents in each block equals the observed number, which is automatic because one of the sufficient statistics is the block by response interaction. This method of fitting our model is efficient because we can avoid distributing nonsample nonrespondents into the household type categories until the last step.

It is possible that with some data sets, some parameters may be inestimable because the maximum likelihood estimates lie on the boundary of the parameter space (are infinite) or because there is no information for the parameter. Inestimable parameters may be removed by reducing the model, but in a production setting it would be unrealistic to attempt to tailor the model specification to each DO, although there might be several versions of the model to use in different types of areas.

If a small amount of prior information is introduced, estimability of all parameters can be guaranteed without the requirement of judgemental intervention in the fitting of each model. A simple prior specification would be given by a prior distribution on all parameters that is normal with mean 0 and a covariance matrix that is diagonal (signifying prior independence) with large variances for all parameters. As long as the variances are large, little bias will be introduced but infinite or inestimable parameters will be pulled toward 0.

An alternative approach to incorporating prior information is to append a small amount of "pseudo-data" to the data set. Because this procedure works directly on the data, rather than the parameter precision matrix, it is applicable when IPF is used. In this approach, we add, to an area, pseudo-data whose proportions by type are equal to those for some surrounding area.

Whatever method is used to estimate model parameters, the next step is to calculate probabilities for each household type in the nonresponse cell for each nonsample block. Using the IPF algorithm, the predictions for the unobserved cells are obtained automatically by applying the same fitting proportions to those cells as to the fully observed part of the table. The estimated counts for each block and household type are then calculated by multiplying predicted proportions by the number of nonresponding addresses in each block.

Once the estimated counts for each block and household type have been calculated, some rounding or imputation procedure must be applied to create a simulated roster. Assuming that an unbiased procedure is used, the choice of rounding procedure affects the variance of the results but not the bias. By an unbiased procedure we mean a stochastic procedure that in expectation imputes the predicted number of units in each cell. Unbiased schemes for

"controlled rounding", i.e. rounding in a two-way table while preserving marginal totals, were developed by Cox (1987) and George and Penny (1987).

Finally, households of the appropriate type must be imputed for the nonsample nonrespondents. There are several possible pools from which to select these donor households. They can be taken from the sampled nonrespondents, the respondents, or a combination of both. Because the work in this paper focuses on evaluating the performance of the loglinear model, we do not discuss rounding and imputation issues any further.

4 Vacant Households

The model and fitting methods described in the previous sections can be applied using any definition of household types. In particular, we could include "vacant" as a household type. Vacant households, however, are a special type of household which does not fit the framework of the model because the distinction between respondent vacants and nonrespondent vacants is not very useful. Respondent vacants are simply those vacant households which were identified as vacant through the mail return of the original questionnaire.

To address this problem, we suggest fitting a separate logistic regression model to first predict the number of nonrespondent vacant households in each block. Then, a loglinear model can be fit to predict the distribution of the non-vacant household types in the remaining nonrespondent households.

5 The structure of the simulations

The primary objectives of this study are to evaluate the bias, variance and MSE of the estimates of demographic aggregates (such as number of households by race, size and tenure), using estimated household compositions for nonresponding addresses in nonsample blocks, at the block, ARA and DO levels.

Answering these questions analytically is not likely to be feasible, given the complexity of the models and sampling scheme and the number of variations of the models that could be examined. Instead, we approach this problem through simulations. The simulations are similar in structure to those described by Schindler (1993) or Isaki, Tsay and Fuller (1994). We focus our attention on the model fitting and prediction steps because these are the steps in which we have proposed innovations relative to Isaki, Tsay and Fuller. Simulations use complete block-level data from the 1990 census. The steps of the simulation are as follows:

1. Blocks are sampled according to the selected sampling scheme and rate.
2. Predicted counts are calculated for each block.
3. Aggregates of interest in the evaluation are calculated based on the predicted counts and compared to values calculated from the complete data.

These steps are repeated enough times to yield adequate estimates of bias, variance, and mean squared error for the target aggregates.

We will compare the performance of our proposed model with two other possible estimation methods. The first alternative estimation approach involves computing the proportion of nonrespondent households in the follow-up sample for the entire DO that are of each type and then using these proportions to impute households in each nonsample block containing nonrespondents. We will refer to this method as the "unstratified ratio method". The second alternative estimation approach involves first stratifying blocks based on some important characteristic, such as race, and then, in each stratum, computing the proportion of nonrespondent households in the follow-up sample that are of each type. Then, in each stratum, we impute households in these proportions for each nonsample block containing nonrespondents. We will call this the "stratified ratio method". This stratified ratio method is a simplified version of the approach taken by Isaki, Tsay and Fuller (1994).

Both of these alternative approaches have advantages and disadvantages. The unstratified ratio method is conceptually simple and easy to carry out. It does not, however, take into account differences between blocks and fails to use any of the respondent information. The stratified ratio method makes more use of the differences between blocks and more use of the respondent information, depending on how stratification is done. However, it fails to take into account intermediate levels of geography. Our method generalizes both of these approaches to an even finer level of detail and so we feel a comparison of these three methods is useful.

In our simulations, we omit the steps of estimating vacant households, rounding, and imputation. Because all three estimation methods would involve these same three steps, removing them allows us to compare the aspects of the methods that differ. Specifically, we removed all the vacant households from the data set before fitting the models. We also did not round the final estimates and we did not do any imputations. We plan to investigate these procedures after we have shown that the prediction model performs well.

Following Isaki, Tsay and Fuller (1994) we classified the non-vacant households into 18 types defined by the cross-classification of households by three size

categories (1-2 people, 3-4 people, 5 or more people), three race categories (non-Hispanic Black, Hispanic, Other), and two tenure categories (owner, renter).

We used short-form data from the 1990 Census for one District Office (DO). Once the 6,771 vacant households were removed, this DO consisted of 4888 blocks with a total of 106,195 households. Of these households, 15.3% were non-Hispanic Black, 6.5% were Hispanic, 78.2% were Other, 32.1% were renters, 67.9% were owners, 53.6% had one to two people, 33.6% has three to four people, 12.7% had five or more people and 77.0% were respondents. The race of a household was determined by the most prevalent race in the household. The data did not contain ARA information, but it did contain block group information. (Block groups are smaller than ARAs.) We formed 94 pseudo-ARAs by grouping consecutive block groups into groups of about 50 blocks. This seemed like a reasonable procedure because block groups close in identification numbers are also geographically close.

We simulated a NRFU sampling procedure with a sampling rate of approximately 30%. To do this for the unstratified ratio method, we drew a sample of 1500 blocks using simple random sampling without replacement from the total number of blocks in the DO. For the stratified ratio method, we stratified the blocks into 59 strata of approximately 83 blocks each based on the racial composition of the blocks, as described in Isaki, Tsay and Fuller (1994). (To do this we used information from both the respondents and nonrespondents, so the blocks were stratified perfectly.) In each stratum, we drew a simple random sample without replacement of 25 blocks. For our method, we drew a 30% sample using simple random sampling without replacement from each "ARA".

For each estimation procedure, the sampling procedure was carried out 30 times and for each sample the model was fit using the information from all the mailback respondents and from the mailback nonrespondents in the NRFU sample, as called for by the particular estimation procedure.

The particular version of the loglinear model that we fit used household type as the x_1 variable, the cross-classification of race by tenure as the x_2 variable, the cross-classification of race by size as the x_3 variable, tenure as the x_4 variable, and an additional $r * a * x_5$ term with race as the x_5 variable.

Before we fit the loglinear model, we did a small amount of empirical Bayes smoothing. This ensured that the model could be fit in every case and also increased the convergence speed of the IPF. We added one respondent household to each block. This household was distributed among the 18 household types according to the overall DO proportions of respondents.

To evaluate the estimates for the nonsample nonrespondents we used the following loss functions. As a measure of the bias for the estimates of the number of households of category j in a geographic unit (e.g. block, ARA, DO) we calculated the Root Mean Weighted Squared Bias. Define the relative error for category j (a type or combination of types) in geographic area i (a block or combination of blocks):

$$d_{ijs} = \frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}}$$

where Y_{ij} is the true number of households of category j in geographical unit i, $\hat{Y}_{ijs}$ is the estimated number of households of type j in geographical unit i using the model fit from sample s, and Y_{i+} is the total number of households in geographical unit i. Then the estimated Mean Weighted Squared Bias is given by

$$\widehat{\text{Bias}}^2 = \frac{\sum_i Y_{i+}\{(\text{Ave}_s(d_{ijs}))^2 - \frac{1}{SY_{i+}^2}\text{Var}_s(d_{ijs})\}}{\sum_i Y_{i+}}$$

where S is the number of samples drawn, $\text{Ave}_s(\cdot)$ is an average over the S samples, $\text{Var}_s(\cdot)$ is a variance of the S samples, and $i = 1, \ldots, I$ where I is the total number of geographical units in the DO. Specifically, $\hat{Y}_{ijs}$ is tabulated as the observed number of households of category j in area i plus the estimated number of nonsample nonrespondent households of category j in area i as predicted by the model fit using sample s. For example, $\hat{Y}_{ijs}$ could be the observed plus estimated number of households of type 3 in block i or it could be the observed plus estimated number of rental households in ARA i. The second term in the numerator removes a bias due to the finiteness of the simulation.

As a measure of the mean square error, we calculated the Root Mean Weighted Mean Squared Error which is given by

$$\widehat{\text{RMSE}}^2 = \frac{\sum_i Y_{i+}\left(\text{Ave}_s(d_{ijs}^2)\right)}{\sum_i Y_{i+}}$$

where Y_{ij}, $\hat{Y}_{ijs}$, Y_{i+}, $\text{Ave}_s(\cdot)$, i, and S are defined as above. We obtain a measure of the standard deviation of the estimates by subtraction:

$$\widehat{\text{SD}} = \sqrt{\widehat{\text{RMSE}}^2 - \widehat{\text{Bias}}^2}.$$

These loss functions were specifically chosen so that measures of error can be calculated at various levels of geography. This reflects the fact that block level estimates are often aggregated to form estimates at higher levels of geography. Therefore, it is important to be able to measure error not only at the block level, but also at these higher levels of

geography. With this in mind, these measures were also chosen because they weight errors by the size of the geographical unit. This leads to consistent estimates of error when aggregating over geographical units. For example, when blocks are weighted by size, two blocks with 5% error will contribute the same amount to the measure of error regardless of whether the blocks are left separate or aggregated into one large block. This is not the case if blocks are not weighted by size. We base our measures on errors relative to the total area i population rather than the population in the target category only, because the latter denominator inflates the importance of small errors in blocks where the category rarely or never appears.

Note that these MSE, bias, and standard deviation measures are all with respect to repeated nonresponse follow-up sampling from the given finite population of blocks.

6 Simulation Results

Some results of the simulation are shown in Figure 1. The nine bar charts in this figure show the weighted mean bias, standard deviation, and RMSE for estimates of the total number of households in each of the tenure categories (only the renter category is shown since the results for the owner category are identical), size categories, and race categories at each of the block, ARA, and DO levels of geography. The height of the bar represents the percent bias, standard deviation, or RMSE and this percent is also printed at the top of each bar. All charts are on the same scale. Estimates for each category were calculated at each level of geography using each of the three methods. The results for each method are represented by the three differently shaded bars, as indicated by the legends.

The results in Figure 1 show that both the stratified ratio method and the loglinear model perform better than the unstratified ratio method by all three measures in almost all cases. Therefore, we will confine our detailed discussion of the results to the comparison of the stratified ratio method and the loglinear model. The loglinear model has less bias, standard deviation and RMSE at all levels of geography for both the tenure and size categories. The difference between the two methods is most dramatic for the tenure categories at the ARA level. The stratified ratio model has slightly larger standard deviation and RMSE for the race categories at all levels of geography and slightly less bias for the race categories at the block and DO levels. The loglinear model, however, shows less bias for the race categories at the ARA level. The results of the stratified ratio method in this simulation may, however, be better than can be expected in practice because blocks were stratified perfectly by race.

7 Future Work

We plan to continue investigating this procedure in several ways. We plan to further examine the use of empirical Bayes smoothing across local areas. We think that "borrowing strength" from neighboring blocks will reduce bias, especially in small blocks, without affecting accuracy at the ARA level.

We also plan to implement the logistic regression model to predict the number of vacant households in each block.

Furthermore, we plan to evaluate the performance of the estimation procedure under various sampling rates and to verify our results using data from other District Offices.

Finally, we plan to investigate extensions of these models to incorporate administrative records in the estimation and imputation phases (Zaslavsky 1995).

References

Cox, L.H. (1987), "A Constructive Procedure for Unbiased Controlled Rounding," *Journal of the American Statistical Association*, 82:520-524.

George, J.A and Penny R.N. (1987), "Initial Experience in Implementing Controlled Rounding for Confidentiality Control," *Proceedings of the 1987 Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, 3:253-262.

Schafer, J.L. (1995) "Model-Based Imputation of Census Short-Form Items," *Proceedings of the 1991 Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 267-299.

Schindler, E. (1993), "Sampling for the Count; Sampling for Non-Mail Returns," unpublished report, Bureau of the Census.

Isaki, C.T., Tsay, J.H. and Fuller, W.A. (1994) "Design and Estimation for Samples of Census Nonresponse," *Proceedings of the 1994 Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 289-305.

Zanutto, E., and Zaslavsky, A.M. (1994) "Models for Imputing Nonsample Households with Sampled Nonresponse Followup," *Proceedings of the Survey Research Methods Section of the American Statistical Association.*, Alexandria, VA.

Zanutto, E., and Zaslavsky, A.M. (1995) "Models for Imputing Nonsample Households with Sampled Nonresponse Followup," *Proceedings of the 1995 Annual Research Conference,* Washington, DC: U.S. Bureau of the Census, pp. 673-686.

Zaslavsky, A.M. (1995), "Estimating a Population Roster from an Incomplete Census Using Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Follow-up," unpublished report, Department of Statistics, Harvard University.

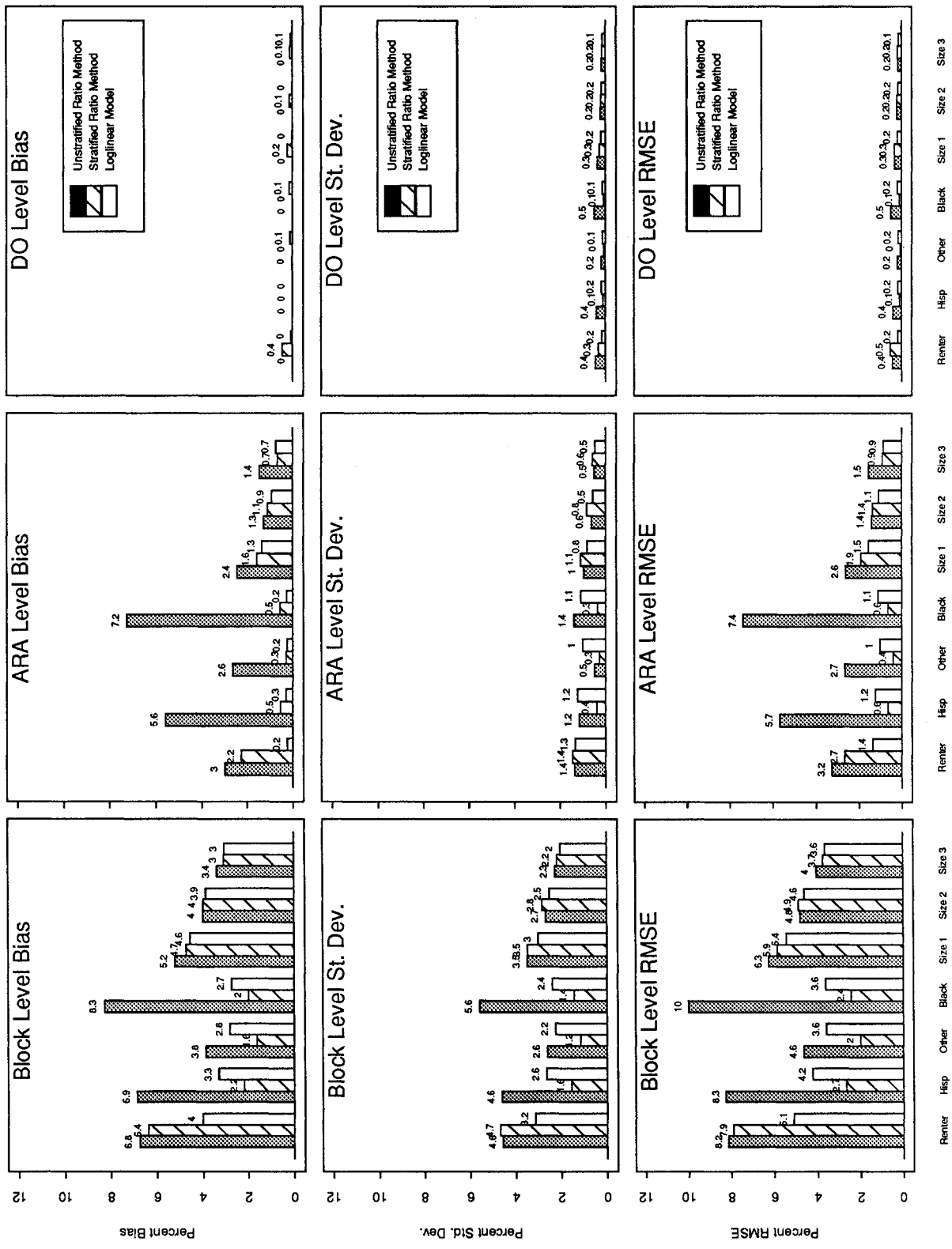

Figure 1: Weighted Average Bias, Standard Deviation, and RMSE at block, "ARA", and DO levels, as a percent of total number of households in each area.