MODELS FOR IMPUTING NONSAMPLE HOUSEHOLDS WITH SAMPLED NONRESPONSE FOLLOWUP

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

Key Words: Nonresponse Followup, Census, Loglinear Models

1 Problem and notation

Sampling for nonresponse followup (NRFU) has been proposed as an 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. Furthermore, there are good reasons for requiring the NRFU sample to be a sample of blocks rather than individual households, having to do with the interaction of NRFU sampling with coverage measurement and the exigencies of field management of NRFU.

The following is a brief description of the 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 followup operation) in an area (say, a District Office). At the second stage, followup (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. Second stage followup is assumed to be complete in the sample blocks, meaning that all addresses either are resolved to be vacant or by 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; one possible household type is "vacant," meaning that no household resided at that address. (Many "vacant" households may have already been detected through mail return of the original questionnaire.)

We assume the following notation:

- i = block index,
- $a = a(i)$ = ARA index corresponding to block i, where the ARA is 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,
- r = first-stage response indicator, $r = 0$ for responding households and $r = 1$ for nonrespondents.

The covariates $x_1, \ldots x_4$ may be multivariate. As a special case, x_1 could be a vector of indicators for membership in each of several classes of households; then the covariate is equivalent to classification of the households into classes. The covariate vectors could also include quantities such as the number of members, number of Black members, or number of Black members of age 18+, for which average values would be meaningful.

The ARA could be replaced by any area intermediate between the DO and the block. Instead of using the ARA as usually defined, areas could be defined by a combination of geographical contiguity (the ARA) and stratification by block-level covariates (such as percent minority), in order to obtain more homogeneous areas whose differences could be described by modeling.

2 A model and its interpretation

We will calculate our imputations by assuming a loglinear model for the following 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 $a = a(i)$.

The interpretations of the terms of the model are stated below. Note that each effect corresponds to a margin of the block×type×response table.

- i Main effect for block, corresponding to the margin for number of households per block (block size);
- x_1 Main effect for household covariates, corresponding to number of households for each type or mean value of each covariate included in x_1;
- r Main effect for response, corresponding to overall level of response across the DO;
- $i * x_2$ Interaction of block and household covariates, corresponding to number of households by type or mean covariate levels, within each block;
- $i * r$ Interaction between block and response, corresponding to overall level of response within each block;
- $r * x_3$ Interaction between response and household covariates, representing the association between household level covariates and the propensity of households of different types to respond;
- $r * a * x_4$ Interaction of response, ARA and household covariates, representing the ARA-specific aspect of the interaction of response and covariates.

In order to understand the motivation for this model, this list of terms and corresponding margins or averages should be considered in light of 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 (fitted, predicted) values for every margin or mean corresponding to an effect in the model are equal to the corresponding observed margins or means. This implies that if we fit the models by maximum likelihood, the (1) fitted block counts, (2) response rates by block, (3) covariate means overall (for x_1 covariates) and (4) by block (for x_2 covariates), and (5) covariate means for nonrespondents overall (for x_3 covariates) and (6) for nonrespondents by ARA (for x_4 covariates) will match those in the observed data. Thus, this model generalizes the model used by Isaki of block×type independence, yielding unbiasedness at smaller levels of aggregation, assuming that the margins and averages are estimated unbiasedly from the data.

We have not yet defined the sets of covariates $x_1, \ldots x_4$ to be included in the model. Because x_4 interacts with the smallest level of aggregation for the nonrespondent data, it should include those covariates which it is most important to impute accurately at the ARA level (and almost as accurately at the block level). These would include numbers of household members by Voting Rights Act category, as well as the indicator for vacant address. ("Vacant" may be considered a household type, for these purposes.) The covariate vectors x_2 and x_3 must include all the components of x_4, at a minimum, but may also include other covariates. Finally, x_1 may be very detailed, including indicators for all observed household types. An interpretation of the way the model treats the different sets of covariates is that we estimate the detailed distribution of household types across the whole area (x_1) and then shift that distribution to allow for the general characteristics of the block (x_2), the general differences between responding and nonresponding households (x_3), and the most important differences between responding and nonresponding households in the particular ARA (x_4).

The model described above can be modified to bring in more or less local detail. For example, we could replace the ARA by a smaller unit, such as a cluster of contiguous blocks containing a single sample block. Or we could leave the ARA in the model, but include the interaction of a lower-dimensional covariate x_5 with the block cluster as just described. Also, see the comments on the choice of "ARA" definitions at the end of the last section. We could simplify the model by omitting interactions, either $i * x_2$ or $r * x_3$ or $r * a * x_4$.

The idea of modeling household characteristics using low-dimensional covariates at the block level and in more detail at more aggregated levels is similar in concept, although not in details, to the model described in Zaslavsky (1992, section 5).

3 Fitting the models and calculating imputations

The model described previously may be formulated in several ways; for each formulation, there is a natural approach to fitting the model. A possible complication in fitting the model is that our data do not form a complete block×response×type table, because we have information on responding households in all blocks but for nonresponding households only in the sample blocks.

The most direct approach to fitting the model is through Poisson regression. The properties of loglinear models that we rely on hold equally well under Poisson and multinomial sampling. The incompleteness of the data present no problem, as the Poisson regression model can be fitted to counts in all available cells.

A second approach is to fit a multinomial logistic

regression model. The cell for this model is defined by block×response (i.e. block and first stage response status are the predictors), and the outcome is household type. Again, the unobserved cells pose no problem because they simply can be omitted from the regression data. An advantage of this approach is that the number of parameters is smaller because the model describes the distribution of household types in each cell but does not describe the marginal distribution of cell counts and response status by cell. The logistic regression model is of the form

$$\log P(j \mid i, r) \sim x_1 + i * x_2 + r * x_3 + r * a * x_4$$

and the parameters corresponding to the effects i, r and $i*r$ do not appear in the model. A disadvantage of this approach is that standard multinomial logistic regression software may be less readily available.

A third approach is to apply the EM algorithm to complete the missing cells and then to fit a loglinear model to the complete table. This approach may be impractical if there are scaled covariates (i.e. if any x is not simply a categorization of household types), because standard loglinear modeling software usually does not handle scaled covariates.

With any of these model formulations, it is possible that with any particular data set, 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 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. Note that this prior information may be incorporated by adding a small amount (the inverse of the prior variances) to the diagonal elements of the information matrix required in the Newton-Raphson algorithms used for fitting these models. A similar procedure for estimation with sparse data was applied by Belin, Diffendal, Mack, Rubin, Schafer, and Zaslavsky (1993). In their application, however, the prior was estimated from the data, while here we concentrate on minimizing bias by keeping the amount of prior information small, even though mean squared error at the block level may not be reduced as much as would be possible with more aggressive smoothing. (An alternative approach to incorporating prior information is to append a small amount of "pseudo-data" to the data set, but given the complexity of the model and the large number of cells, this approach may be less convenient in this setting.)

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. These probabilities are predicted directly from the multinomial logistic regression model. The Poisson regression model predicts counts for these cells (up to a proportionality constant determined by the unestimated block×response effect for that cell), which can be turned into predicted proportions. Similarly, predicted cell counts from the loglinear model may be turned into predicted proportions. 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. Note that the margins by block are integers, but the margins by household type are generally not integers.

Finally, 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. The simplest unbiased procedure is simply to impute households independently, one by one, according to the predicted probabilities in each block. Variance can be reduced by attempting to control the number of households by class, or other aggregates such as the total number of Blacks aged 18+, to be close to the predicted number. 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).

4 The structure of the simulations

The primary objectives of the study are to evaluate the bias, variance and MSE of the estimates of demographic aggregates (such as number of persons by race and age), using imputed household compositions for nonresponding addresses in nonsample blocks, at the block, ARA and DO levels. We want

to know whether we can attain acceptable levels of error under various sampling schemes and rates, and to try to determine the best models and sampling schemes.

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 will be examined. Instead, we approach this problem through simulations. The simulations are similar in structure to those described by Schindler (1993). The steps of the simulation are as follows:

1. Blocks are sampled according to the selected sampling scheme and rate.
2. A model is fitted as described above.
3. Predicted counts are calculated for each block; the aggregates of interest in the evaluation are calculated based on the predicted counts.
4. Counts are rounded, if this is part of the methodology, and households are imputed, for each block; the aggregates of interest in the evaluation are recalculated based on the imputations or rounded counts.

Steps 1 through 4 are repeated enough times to yield adequate estimates of bias, variance, and mean squared error. Errors are estimated for aggregates at both steps 3 and 4. The purpose of calculating error at both steps is to estimate the contribution of rounding and imputation to error; if error is at acceptable levels at step 3 but is much larger at step 4, it would make sense to retain or even simplify the prediction model and concentrate on improving the imputation procedures.

There are at least two simulation parameters whose effect we would like to investigate. These are the NRFU sampling rate (fraction of blocks sampled) and the truncation point for nonsample data collection (i.e. limited to mailback, or including some fraction of NRFU cases; a June 2 cutoff date, with sampling thereafter, would be interesting to look at). There may be other design features that could be investigated. It would probably be worthwhile to run the simulations at several levels of these factors, at least in a few DOs.

5 Preliminary Simulation Results

As a preliminary approach to our simulations, we restricted ourselves to steps 1-3 of the simulation procedure described in the previous section. This allowed us to evaluate the performance of the prediction model without the added contribution to error by the rounding and imputation. We plan to investigate rounding and imputation procedures after we have shown that the prediction model performs well.

Following Isaki, Tsay and Fuller (1994) we classified households into 19 types where 18 of these types are defined by the cross-classification of households by three size categories, three race categories, and two tenure categories. The three size categories are one to two people, three to four people, and five or more people. The three race categories are non-Hispanic Black, Hispanic, and Other. The two tenure categories are owner and renter. The 19th type is for vacant households.

We used short-form data from the 1990 Census for one District Office (DO). This DO consisted of 4907 blocks with a total of 112,966 households. Of these households 14.38% were non-Hispanic Black, 6.13% were Hispanic, 73.50% were Other, 30.20% were renters, 5.99% were vacant and 72.60% were respondents. The race of a household was determined by the most prevalent race in the household. (Only 2.29% of the households had more than one race present in the household). This data did not contain ARA information so we divided the blocks into 10 pseudo-ARAs of approximately 491 blocks each, based on block identification numbers. This seemed like a reasonable procedure because blocks close in identification numbers are also geographically close.

To simulate a NRFU sampling procedure with a sampling rate of 30%, we drew 15 samples of 1472 blocks each, using simple random sampling without replacement from the total number of blocks in the DO. For each sample, the model was fit using the information from all the mailback respondents and from the mailback nonrespondents in the NRFU sample. We fit one specific version of the model with household type as the x_1 variable, the cross-classification of race by tenure (plus the vacant category) as the x_2 variable, the cross-classification of race by size (plus the vacant category) as the x_3 variable, and tenure as the x_4 variable. To fit this model, we used a combination of Iterative Proportional Fitting and the EM algorithm (i.e. the third method of Section 3).

To evaluate the estimates for the nonsample nonrespondents we used two loss functions. As a measure of the bias for the estimates of the number of households of type j in a geographic unit (e.g. block, ARA, DO) we calculated the Root Mean Weighted

Squared Bias which is given by

$$\hat{B}_j^* = \sqrt{\frac{\sum_i Y_{i+}(\hat{B}_{ijs}^2 - \hat{V}_{ij})}{\sum_i Y_{i+}}}$$

where

$$\hat{B}_{ijs} = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}}\right)$$

and

$$\hat{V}_{ij} = \frac{1}{SY_{i+}^2}\left(\frac{\sum_{s=1}^{S}(\hat{Y}_{ijs} - \overline{\hat{Y}}_{ij})^2}{S-1}\right)$$

and Y_{ij} is the true number of households of type 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, Y_{i+} is the total number of households in geographical unit i, S is the number of samples drawn, $\overline{\hat{Y}}_{ij}$ is an average over the S samples, and $i = 1, \ldots, N$ where N is the total number of geographical units in the DO. Specifically, $\hat{Y}_{ijs}$ is calculated as the observed number of households of type j in area i plus the estimated number of nonsample nonrespondent households of type 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. As a measure of the mean square error, we calculated the Root Mean Weighted Squared Root Mean Squared Error which is given by

$$\hat{E}_j^* = \sqrt{\frac{\sum_i Y_{i+}\hat{E}_{ij}}{\sum_i Y_{i+}}}$$

where

$$\hat{E}_{ij} = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}}\right)^2$$

where Y_{ij}, $\hat{Y}_{ijs}$, Y_{i+}, i, and S are defined as above.

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.

Some results of the simulation are shown in Tables 1 and 2. In both of these tables, the columns indicate the type of household that was estimated. The rows indicate the bias and root mean square error measures, as defined above, at various levels of aggregation (block, ARA, DO). The last row of each table lists the prevalance of that particular type in the DO. All entries in the tables are percentages.

Both tables show that the model predicts the missing data very well. They show small errors at the block level and very small errors at the ARA and DO levels. In fact, in most cases the errors are smaller than 1%, which is a smaller amount of error than we would expect to get as a result of undercount.

The results for predicting the number of households of types 1-9 are shown in Table 1. The results for types 10-19 were similar. To briefly summarize the results for all 19 types (not all shown in Table 1), when predicting the number of households at the block level, the biases ranged from 0.40% to 2.71% with an average of 0.89% and the root mean squared errors ranged from 0.64% to 5.12% with an average of 1.48%. At the ARA level, the biases ranged from 0.01% to 1.05% with an average of 0.13% and the root mean squared errors ranged from 0.02% to 1.76% with an average of 0.24%. At the DO level, the biases ranged from 0.00% to 0.87% with an average of 0.09% and the root mean squared errors ranged from 0.01% to 1.39% with an average of 0.15%. Results for predicting the number of households of a particular race or tenure are shown in Table 3. In this table, biases at the block level ranged from 1.27% to 2.71% with an average of 2.11% and root mean squared errors ranged from 2.25% to 5.12% with an average of 3.83%. Biases at the ARA level ranged from 0.19% to 1.05% with an average of 0.54% and root mean squared errors ranged from 0.33% to 1.76% with an average of 0.94%. At the DO level the biases ranged from 0.10% to 0.87% with an average of 0.43% and the root mean squared errors ranged from 0.18% to 1.39% with an average of 0.70%.

These tables show that the relative error decreases as we aggregate over larger levels of geography. This is a characteristic of the model because the model includes more information about the households at higher levels of geography. These tables also show that the contribution to error due to bias and due to variance are of approximately the same magnitude at all levels of aggregation.

	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8	Type 9
Block Bias	0.40	0.53	0.77	1.11	1.41	1.65	0.44	0.62	0.79
Block RMSE	0.64	0.89	1.28	1.88	2.30	2.83	0.72	1.04	1.31
ARA Bias	0.01	0.07	0.05	0.30	0.09	0.17	0.03	0.05	0.12
ARA RMSE	0.02	0.13	0.09	0.49	0.19	0.36	0.05	0.10	0.21
DO Bias	0.00	0.02	0.02	0.14	0.03	0.12	0.01	0.02	0.06
DO RMSE	0.01	0.04	0.04	0.23	0.09	0.22	0.02	0.05	0.10
Prevalence	0.75	1.21	3.01	3.95	27.12	14.38	0.94	1.35	2.50

Table 1: Household Types 1-9

	Black	Hispanic	Other	Owner	Renter	Vacant
Block Bias	1.87	1.27	2.09	2.04	2.68	2.71
Block RMSE	3.36	2.25	3.81	3.61	4.80	5.12
ARA Bias	0.57	0.19	0.39	0.40	0.65	1.05
ARA RMSE	0.92	0.33	0.75	0.69	1.17	1.76
DO Bias	0.35	0.10	0.40	0.33	0.54	0.87
DO RMSE	0.57	0.18	0.65	0.52	0.87	1.39
Prevalence	14.38	6.13	73.50	63.80	30.20	5.99

Table 2: Race and Tenure

6 Future Work

We plan to continue investigating this procedure by evaluating the specific model described in the previous section with many more samples. We also plan to evaluate other versions of the model as well as the effect of different sampling rates and different types of households classifications. We also plan to examine the use of Empirical Bayes Smoothing across local areas. This would indicate how much "borrowing strength" from neighboring blocks can improve our estimations. Finally, we also plan to look at the rounding and imputation stages of the procedure.

References

Belin, T.R., Diffendal, G.J., Mack, S., Rubin, D.B., Schafer, J.L., and Zaslavsky, A.M. (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," *Journal of the American Statistical Association*, 88:1149-1166.

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. "Initial Experience in Implementing Controlled Rounding for Confidentiality Control," *Proceedings of the Bureau of the Census Annual Research Conference*, 3:253-262.

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 Bureau of the Census Annual Research Conference.*

Zaslavsky, A.M. (1992), "Representing the Undercount by Multiple Imputation of Households," unpublished manuscript, Department of Statistics, Harvard University.