Potential Applications of Model-Assisted Estimation to Demographic Surveys in the U.S.

Robert E. Fay[1]

U.S. Census Bureau, 4700 Silver Hill Road, Washington, DC 20233-9001, robert.e.fay.iii@census.gov

Abstract

For decades, the U.S. Census Bureau has conducted a number of large-scale household surveys, including the venerable Current Population Survey. The American Community Survey (ACS) recently became the largest. Until the ACS, samples for these surveys were usually drawn using a mixture of address and sometimes area frames from the previous decennial census, supplemented by frames for new construction. Although census data were reflected in the sampling design, estimation for the ACS and other surveys has been based on complex forms of ratio estimation to independent totals, without further incorporating frame information.

Beginning with Census 2000, the U.S. Census Bureau now maintains a Master Address File (MAF) as a current inventory of housing units. The MAF provides the sampling frame for the ACS and the redesigned versions of the other major demographic surveys. This change to a housing-unit frame presents an opportunity to investigate possible model-assisted approaches to estimation. Model-assisted estimation affords an approach to incorporate information from the frame into the estimation, information that might include data from the previous census or administrative records. Moreover, model-assisted estimation uses auxiliary information in an essentially design-unbiased way, whereas model-based alternatives generally share the risk of introducing substantial bias. This paper will report on initial results for ACS using data for 36 test counties during 1999-2001. It will also suggest how similar principles could be applied to other demographic surveys.

1. Introduction

This paper suggests a new look at an old problem: estimation strategies for large-scale demographic surveys in the U.S. The immediate scope is somewhat narrower—demographic surveys in the U.S. in which the Census Bureau has drawn the sample and retains information about the frame—but even this narrowed scope encompasses important components of the U.S. statistical system, including the American Community Survey (ACS), the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), the American Housing Survey (AHS), and several others.

For years—and in most cases, decades—*raking/ratio estimation,* originating with the 1940 paper of Deming and Stephan, has been the backbone of the estimation strategy for each of these surveys. Similarly, estimation for the U.S. decennial long-form sample has also been accomplished with this method for several censuses, including Census 2000. Generally, raking/ratio estimation proportionally adjusts survey weights to achieve consistency between the survey estimates and independent controls. To provide additional background, Section 2 of this paper remarks on the strengths of raking/ratio estimation and typical uses in major U.S. Census Bureau surveys.

Alternatives to raking/ratio estimation have also been studied in the statistical literature. Section 3 reviews some of the critical references, called here *the primary thread,* in the literature on generalized regression estimation, calibration estimation, and related estimators. Section 4 reviews a somewhat separate part of the literature, *the American thread,* centered on U.S. applications, particularly for government statistics.

With the exception of some applications noted in section 4, the influence of these methods on U.S. government statistics has thus far been relatively limited, especially with respect to the surveys reviewed in Section 2. Section 5 describes the goals for the ACS to produce small domain estimates at the tract and block-group level, and Section 6 argues how the combination of administrative records and model-assisted estimation might help to meet these goals. Section 7 suggests that the ACS application might not represent the only opportunity of its kind—that changes in sampling frames and refinements in the statistical use of administrative records are increasing the potential application of generalized regression estimation and other model-assisted estimators.

The ACS has been developed through several years of theoretical work and empirical testing. Full-scale implementation of the ACS began in January of 2005, with a monthly sampling rate of approximately 1-in-480, yielding an annual sample of approximately 1-in-40. The ACS has become the largest demographic sample survey in the U.S. statistical system.

Prior to 2005, tests have included samples in 36 counties at a sampling rate approximating the full-implementation rate (34 of the 36 counties were sampled at 3% or 5% per year, beginning in 1999). In 2000, a national test was also conducted, the Census 2000 Supplemental Survey.

Based on an ongoing, fully-implemented ACS, the 2010 Census will contain neither a long-form nor a long-form sample. For the ACS to replace the decennial long-form, which in Census 2000 was based on approximately a 1-in-6 sample, the 1-in-40 annual ACS sample will be accumulated over a period of 5 years to provide a 1-in-8 sample of households for use for small geographic areas, such as census tracts or block groups.

Thus far, ACS estimation has depended on raking/ratio estimation, and no departure from this strategy is envisioned for estimation from the 1-year data for 2005. As section 5 will detail, however, evidence from the 36 test counties suggests that an alternative approach might be needed to produce tract-level estimates from the 5-year samples. Preliminary research, reported in a paper presented at the Joint Statistical Meetings (JSM) (Fay 2005), suggests that model-assisted estimation combined with data from administrative records could provide substantial improvements in the small area ACS estimates. Although the research is far from complete, the preliminary results appear promising, and Section 6 summarizes some of them.

The final section comments on possible extensions of this research to other surveys. Successful application to the ACS will help to guide future research, of course, but even should the methods encounter a practical obstacle blocking implementation to ACS, it remains possible that model-assisted estimation could be adapted to one or more other demographic surveys.

2. Features of Raking/Ratio Estimation in Demographic Surveys

Ratio estimation is a standard topic of textbooks on survey design. It is also one of the most practically useful tools in estimation. Ratio estimation can be used to achieve consistency between a weighted survey estimate and a fixed total or disjoint set of fixed totals. Raking/ratio estimation typically can achieve consistency with more than one marginal total simultaneously. The margins must be feasible, that is, consistent with some possible table. (Convergence can sometimes be affected by the presence of zero cells.) Raking/ratio estimation is typically achieved through proportional adjustment of survey weights iteratively, by cycling through sets of marginal constraints.

In application, raking/ratio estimation often achieves three purposes:

1. Achieving consistency with independent estimates
2. Reducing the variance of other characteristics
3. Reducing bias

In some applications, bias reduction is the most important objective. In particular, raking/ratio estimation may partially compensate for deficiencies in the sampling frame, differential nonresponse, coverage error, or other measurement errors. But the presence of bias in the control totals also generally results in inducing bias into the adjusted estimates. Whether raking/ratio estimation actually reduces bias in a particular application depends in large part on the relative degree to which the bias in the unadjusted estimates exceeds that of the controls.

The Current Population Survey: For example, monthly estimation in CPS includes several stages, including a stage to incorporate composite estimation into the weights. Raking/ratio estimation is used to achieve consistency with population estimates, both detailed national estimates by age, sex, race, and ethnicity, and for coarser estimates by state. The associated Annual Social and Economic Supplement includes raking/ratio estimation with additional stages of estimation. In general, raking/ratio estimation achieves

1. Consistency with independent estimates of the civilian, non-institutional population. (The most readily available published population estimates, however, are for the total population rather than the civilian, non-institutional population).

2. Reduction in variance (especially for estimated totals).

3. Reduction in bias, especially bias due to undercoverage of some demographic groups such as adult Black males and, to a lesser extent, males generally.

For the most part, estimation for the CPS and for many other demographic surveys makes minimal use of information from the frame at the unit level. In past decades, information from the frame appeared to offer little except an accurate account of the probability of selection. In the CPS, the ultimate sampling unit was originally not an individual housing unit but a compact cluster of 4 housing units. Until recently, matching the CPS to any other survey or census was a complex undertaking. Thus, until now practical circumstances have discouraged a notion of using frame characteristics to improve the CPS or other federal surveys.

The Decennial Census Long-Form: Through Census 2000, raking/ratio estimation has been used to adjust census long-form weights to conform to control totals from the 100% count. In general, this approach achieved

1. Consistency with 100% count. Raking/ratio estimation was implemented separately in each weighting area, an area that often coincided with a census tract. As a general requirement, weighting areas were required to have at least 400 sample people, so areas were combined until this goal was achieved.

2. Reduction in variance, especially for estimated totals.

3. A reduction in bias from differential nonresponse in the long-form sample (for example, some highly incomplete long-forms were converted to short-form cases).

Arguably, estimation for the census long-form uses information from the frame—the 100% items from the short form. In other respects, however, estimation for the decennial census resembles estimation for the CPS, so that it is possible to view estimation in the census as controlling to independent totals rather than to recognize that it uses information from the frame.

The ACS as currently planned: Current procedures employ raking/ratio estimation, with county-level population estimates used as controls. The objectives will be similar to those of the CPS, but carried out in more geographic detail.

1. Consistency with independent population estimates for the ACS universe.

2. Reduction in variance (especially for totals).

3. Reduction in bias, if some demographic groups, such as adult Black males, are not covered in ACS as well as other groups.

There are currently no suitable tract-level population controls available for the ACS. Without tract-level weighting areas, the preliminary estimates in the test counties during 1999-2001 exhibited high variances. Ongoing research to be described here is investigating whether one or more stages of model-assisted estimation could be incorporated into the weighting to improve small-area estimates. The proposal would retain raking/ratio estimation at higher levels in the final stages of ACS weighting. The objectives of the stages of model-assisted estimation would

1. Achieve consistency or close agreement with some totals based on administrative records. (However, this would be a secondary consequence, not the primary objective, of the modification.)

2. Reduce the sampling variance of the small-area estimates, particularly of totals, without adversely affecting the reliability of higher-level estimates.

3. Have negligible effect on the bias.

The third objective, the avoidance of appreciable bias, is a feature of model-assisted estimation under appropriate conditions. The next two sections review some of the extensive literature on the subject.

3. Generalized Regression and Calibration Estimation – Primary Literature Thread

Särndal, Swensson, and Wretman's text (1992) presents the general category of *model-assisted estimators* and examines specific important classes of these estimators, including *generalized regression estimators.* The general category of model-assisted estimators can be motivated by an appeal to models to incorporate auxiliary information. A familiar member of the category, ratio estimators, illustrates general properties of the rest of the category: although not always strictly design unbiased, the bias of model-assisted estimators is of lower order than the standard error. Consequently, model-assisted estimators are nearly design unbiased.

Consider the estimation of a population total $\hat{Y}$ for a population with values $y_1,\ldots,y_N$ based on a sample s drawn according to probabilities π_k. For each element k, let there be a vector or auxiliary data x_k of dimension J, with elements x_{jk}; the complete matrix of auxiliary data can be expressed as $X = [x_{jk}]$. Let $W_k^{(0)} = \pi_k^{-1}$, $\hat{Y}^{(0)} = diag(W^{(0)})y$, $\hat{Y}^{(0)'} 1_n = \sum_s y_k / \pi_k$, and $\hat{X}^{(0)} = x\, diag\,(W^{(0)}) = [W_k^{(0)} x_{jk}]$. The regression estimator may be written

$$\hat{Y}_{rg} = \hat{Y}^{(0)'} 1_n + \hat{B}'(X1_N - \hat{X}^{(0)}1_n) \qquad (1)$$

where

$$\hat{B} = (\hat{B}_1,...,\hat{B}_J)' = \left(\sum_s x_k x_k' / \sigma_k^2 \pi_k \right)^{-1} \sum_s x_k y_k / \sigma_k^2 \pi_k \qquad (2)$$

Estimator (2) is based on a model ξ for the underlying population, where each y_k is the realization from a random variable Y_k with expected value $E_\xi(Y_k) = \sum_{j=1}^{J} \beta_j x_{jk}$, and variance σ_k^2. Equation (2) accounts for the joint roles of the model (through σ_k^2) and design probabilities π_k in estimating the regression.

Eq. (1) expresses the regression estimator as the Horwitz-Thompson estimator plus a regression-based correction for the difference between the population totals for the auxiliary data and the weighted estimate from the observed sample. An alternative expression is

$$\hat{Y}_{rg} = \hat{B}' X 1_N + \hat{e}^{(0)} 1_n \qquad (3)$$

where $\hat{e}^{(0)} = [W_k^{(0)} e_k]$ is a 1 x n vector of weighted residuals, $e_k = y_k - \hat{B}' x_k$. Eq. (3) shows the estimator as the regression prediction for the population adjusted by a correction term equal to the weighted sum of residuals from the regression.

Eq. (1) - (3) are expressed in terms of a chosen characteristic, Y. A well-known (e.g., Särndal, Swensson, and Wretman 1992, p. 232) relationship shows that the regression estimator can be re-expressed in terms of an adjustment, g_{ks}, to the initial weights $W_k^{(0)}$.

$$g_{ks} = 1 + (X1_N - \hat{X}^{(0)}1_n)' \left(\sum_s x_k x_k' / \sigma_k^2 \pi_k \right)^{-1} x_k / \sigma_k^2 \qquad (4)$$

giving $W_k = g_{ks} W_k^{(0)}$. Eq. (4) shows that the adjustment does not depend on Y.

Regression estimators are often members of a larger class of *calibration estimators,* in the sense of Deville and Särndal (1992). Subsequent literature on regression estimation, including Rao (1994) and Fuller (2002), notes this connection. For example, Bankier and colleagues (1992, 1996, 2003) motivate the regression estimator for the Canadian census sample data in part through the general framework of calibration estimation. For simplicity, this paper will limit its discussion to potential applications of regression estimation to illustrate the potential of model-assisted estimation. Some situations may merit closer examination of alternative calibration estimators.

The Canadian work and the notation used by Bankier and colleagues was summarized in some detail in (Fay 2005), because it is both a highly notable application and one that parallels many of the issues faced in the ACS application.

4. Applications of Regression Estimation – The American Thread

Although the primary purpose of this paper is to raise awareness of regression estimation and other forms of model-assisted estimation, the issue has also attracted the interest of American statisticians over several years. I have attempted to locate some of this work. For whatever reasons, this thread is less acknowledged in the primary statistical literature. For example out of the work referenced in this section, Fuller (2002) cites only Fuller, Loughin, and Baker (1994). It is appropriate to acknowledge this previous work here, however, because some of it anticipates the point I am making here. The American thread should also be of interest to those like me attempting to apply the theory to government statistics in the U.S.

Some authors advanced aspects of the theory. Early papers of Luery (1986) and of Alexander (1987) anticipated the more complete treatment of calibration estimation by Deville and Särndal (1992). Alexander's treatment is notable for discussing principal person weighting, a common technique in U.S. Census Bureau household surveys to define a household weight, along with alternatives such as generalized regression estimation. Alexander correctly pointed out that measurement error and survey undercoverage affect the properties of estimators. (These considerations are not present, however, when the auxiliary data for both the sample and population are derived from a single source, as is the case in the ACS application to be discussed here.) Jayasuriya and Valliant (1995, 1996) further compared properties of regression estimation, calibration estimation, and principal person weighting in the context of the Consumer Expenditure Survey. The authors investigated a form, restricted regression estimation, that maintained positive weights.

Other work emphasizes applications. Zieschang (1990) described a notable use of generalized regression estimation for the Consumer Expenditure Survey in the U.S. Fuller, Loughin, and Baker (1994) applied regression estimation to the National Food Consumption Survey, primarily as an approach to nonresponse. Their paper, citing previous applications, also reported their approach to achieve positive weights. Kaufman and Scheuren (1995) presented preliminary work to apply regression estimation to NCES surveys, and Scheuren (1996) briefly remarked on this research direction.

Among this work, there has been little emphasis on the potential use of regression estimation or other forms of model-assisted estimation for small domain estimation. Over 20 years ago, Särndal (1984) visited this issue, which is also the central aspect of the potential application to ACS to be described in the next section.

5. Small area estimation in the ACS

Long-form estimates from Census 2000 are published for a number of types of geographic areas, including tracts and block groups. Although tracts vary in size, their ideal size is approximately 4000 people, and that of a block group is 1500. There is a close, although not exact, correspondence between weighting areas in Census 2000 and tracts. The block group is generally the smallest level of long-form publication. Because raking/ratio estimation considered tract but not block-group level controls, it is likely that block-level estimates do not receive as much benefit, if any, from the raking/ratio estimation, in compared to the reduction in variance achieved for tract-level estimates.

Out of the 36 ACS test counties in 1999-2001, 34 were sampled at either 3% or 5% per year, yielding approximately 9% or 15% over three years. The 34 provide a suitable basis to investigate tract-level estimation for the full-scale ACS, which will yield approximately a 12.5% sample over 5 years. The sampling rate in Census 2000 was roughly 17% overall. (Two of the 36 counties were sampled at 1%, yielding a much smaller sample at the end of the 3-year test period. These two counties can be omitted from analyses of small area estimation.)

When differences in sample sizes are taken into account in the 1991-2001 test counties, variances for county-level ACS estimates approximately correspond to Census 2000 variances. At the tract level, however, ACS variances tend to be too large (Starsinic 2005, Fay 2005). In hindsight, the absence of tract-level controls from ACS weighting accounted for the high tract-level variances.

Direct imitation of decennial estimation suggests the approach of attempting to obtain tract-level controls for ACS estimation. The Census Bureau does not now have a program to produce post-censal population estimates to this level of geographic detail. Although in principle such an approach might succeed, an assessment of the level of error in the

hypothetical tract-level estimates would be a necessary requirement to measure tract-level accuracy of the new estimates. This difficulty would occur for tract-level estimates based specifically on administrative record data, for example.

Although his research is at too preliminary a stage to summarize it further, Donald Malec is investigating an approach requiring matching of administrative record data at the person level.

Possible ACS tract-level estimation: An approach to be described here and detailed further elsewhere (Fay 2005) combines matching administrative record information to the ACS frame (the Master Address File or MAF), and using the resulting housing unit information in a model-assisted manner to reduce small-area variances. The proposal would retain raking/ratio estimation at higher levels as the final stages of ACS weighting. As previously noted, the objectives of the stages of model-assisted estimation would

1. Probably achieve consistency with some totals based on administrative records, but this would be a secondary consequence, not the primary objective.

2. Reduce the sampling variance of the small-area estimates, particularly of totals, without adversely affecting the reliability of higher-level estimates.

Model-assisted estimation would not address the third objective, reduction of bias, because the model-assisted estimation will not alter the expected values of the estimates in any systematic manner. Conversely, the model-assisted estimation will not introduce new components of bias that were not previously present. Thus, the success of the effort can be objectively measured by variance calculations.

In one respect, model-assisted estimation may help the ACS improve on an aspect of decennial estimation: variances for block group estimates. This goal will be further discussed in section 6.

6. The Potential for Model-Assisted Estimation in ACS

The goal of the research is to integrate administrative record data into ACS estimation. It should be noted that the administrative data extracted for my research use does not contain names, Social Security numbers, or IRS data. The extracted data includes records for individual persons with age, sex, an assigned race compiled from different sources, and an assigned Spanish origin from different sources. Persons are grouped into households by the MAF identifier. Administrative record data that could not be assigned to a MAF identifier are excluded. For purposes of comparison, 100% data from Census 2000 are also available. The initial study is based only on administrative record data for year 2000. As the results below suggest, full-scale applications could benefit from the use of administrative records available for each of the years in the ACS reference period. Further details are given in Fay (2005).

The regression estimator (1) will result in a reduction in the sampling variance of the estimated total to the degree that the regression successfully predicts the unit-level values of the characteristic, Y. In the extreme, the regression estimate for a characteristic completely fit by the auxiliary data will have no variance. As an initial diagnostic step, a set of unweighted regressions has been fitted to the ACS data for total population. Again, further details are provided elsewhere (Fay 2005).

For simplicity, only two regression models will be presented here. The first predicts total persons in the ACS household (including 0 in vacant units) either from the 3 variables:

1. An intercept term,
2. An indicator if occupied in Census 2000,
3. The number of persons in Census 2000,

or from

1. An intercept term,
2. An indicator if any administrative records persons matched to the MAF,
3. The number of administrative record persons matched to the MAF.

Table 1 presents the preliminary findings.

Table 1 R^2 values from unweighted regression fit at the housing unit level to predict the number of ACS persons. The regression equations were fit to data from 36 ACS test counties, 1999-2001.

	3-var. census	3-var. admin rec
1999-2001	.554	.475
1999	.536	.485
2000	.636	.522
2001	.492	.416

Not surprisingly, the census data predict most successfully in the census year. Reduced but still respectable predictions are available one year away. Similarly, the administrative record data for 2000 predict the number of ACS persons best for 2000. Although not making quite so strong a prediction as the census, the results from the administrative record data suggest a possible reduction in the variance of estimated persons at the tract level of approximately 50%.

These results are naturally preliminary. The next steps include attempting to compute the g-weights (4) at the tract level and evaluating their variance impact through replication. Tract-level estimates are the initial goal. Estimation for the Canadian Censuses in 1991, 1996, and 2001 has employed a two-step method that could be adapted to the ACS situation to achieve both lower block-group variances and the tract-level objectives.

7. Other Potential Applications of Model-Assisted Estimation

As noted, the usefulness of model-assisted estimation for small domain estimation has received comparatively light attention in the American thread, but this feature drives both its use in the Canadian census and its potential usefulness for ACS. Although the primary objectives of most surveys are national estimates, the CPS and other surveys also produce estimates for subnational geographic areas.

For ACS tract-level estimates, there are currently no available model-based population controls to compete with the suggested model-assisted approach. At higher levels of geography, however, the Census Bureau's program of post-censal population estimation may produce alternative estimates. The application of model-assisted estimation in the manner suggested here, which could reduce variance while not appreciably affecting bias, could provide, for purposes of evaluation, improved direct estimates to compare to model-based approaches.

The administrative data used here provide only basic demographic characteristics, and they are protected under Title 13 of the Census Bureau's enabling legislation. Characteristics with greater sensitivity, such as income, are subject to even higher levels of restriction. Theoretically, data on income or other sensitive administrative characteristics potentially could improve the estimation of other characteristics, but only if an appropriate approach to policy questions can be found.

Note: [1] This report is released to inform ongoing parties of ongoing research and to encourage discussion of work in progress.

References

Alexander, C.H. (1987), "A Class of Methods for Using Person Controls in Household Weighting," *Survey Methodology, 13,* 183-188.

Bankier, M.D., Rathwell, S., and Majkowski, M. (1992), "Two Step Generalized Least Squares Estimation in the 1991 Canadian Census," *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 764-769.

Bankier, M., Houle, A.-M., and Luc, M. (1997), "Calibration Estimation in the 1991 and 1996 Canadian Censuses," *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 66-75.

Bankier, M. and Janes, D. (2003), "Regression Estimation of the 2001 Canadian Census," *Proceedings of the 2003 Joint Statistical Meetings on CD-ROM,* American Statistical Association, Alexandria, VA, pp. 442-449.

Deville, J. and Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association, 87,* 376-382.

Fay, R.E. (2005), "Model-Assisted Estimation for the American Community Survey," presented at the Joint Statistical Meetings, Minneapolis, MN, 7-11 August, 2005.

Fuller, W.A. (2002), "Regression Estimation for Survey Samples," *Survey Methodology, 28,* 5-23.

Fuller, W.A., Loughin, M.M., and Baker, H.D. (1994), "Regression Weighting in the Presence of Nonresponse with Application to the 1987-1988 Nationwide Food Consumption Survey," *Survey Methodology, 20,* 75-85.

Jayasuriya, B. and Valliant, R. (1995), "An Application of Regression and Calibration Estimation to Post-Stratification in a Household Survey," *Proceedings of the Survey Research Methods Section,* American Statistical Association, pp. 902-907.

_____ (1996), "An Application of Restricted Regression Estimation in a Household Survey," *Survey Methodology, 22,* 127-137.

Kaufman, S., Li, B., and Scheuren, F. (1995), "Improved GLS Estimation in NCES Surveys," *Proceedings of the Survey Research Methods Section,* American Statistical Association, pp. 149-154.

Luery, D.M. (1986), "Weighting Sample Survey Data Under Linear Constraints on the Weights," Proceedings of the Social Statistics Section, American Statistical Association, pp. 325-330.

Särndal, C.-E. (1984), "Design-Consistent Versus Model-Dependent Estimation for Small Domains," *Journal of the American Statistical Association, 79,* 624-631.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling,* Springer-Verlag, New York, NY.

Scheuren, F. (1996), "Administrative Record Opportunities in Educational Survey Research," in G. Hoachlander, J.E. Griffith, and J.H. Ralph (eds.) *From Data to Information: New Directions for Educational Statistics,* National Center for Educational Statistics, Department of Education, NCES 96-901.

Starsinic, M. (2005), "Comparison of American Community Survey and Census 2000 Long Form Variance Estimates," paper to be presented at the Joint Statistical Meetings, Minneapolis, MN, 7-11 August 2005.

Zieschang, K.D. (1990), "Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey," *Journal of the American Statistical Association, 85,* 986-1001.