Estimation Using Multiple Surveys

Sharon L. Lohr
Arizona State University, Department of Mathematics
Box 1804
Tempe, AZ 85287-1804 USA
sharon.lohr@asu.edu

1. Introduction

Sample survey theory may be thought of as beginning approximately 100 years ago, with A. N. Kiaer's advocacy of representative sampling rather than complete enumeration. At that time, and through the middle of the 20th century, interest centered on official statistics at the national level — estimating the total number of unemployed men in the United States or the total jute production in India. Beginning in the 1960's, however, increasing attention has been paid to other uses of survey data. Small area estimation techniques employ models to attempt to improve the precision of estimates for domains with insufficient sample size. Theory has been developed for using multiple regression, analyzing contingency tables, and performing multivariate analyses with data from complex surveys, thus allowing survey data to be used to answer questions of scientific and sociological interest that are not primary goals of the survey. Longitudinal surveys are increasingly prevalent, and allow measurement of changes in individuals over time. In all arenas, the demands on survey data have increased.

While one survey may be used for providing snapshots of unemployment at the national level, finer summaries and multivariate relations often require more information than a single survey can provide. With data available from many different surveys, it would be cost-efficient to pool the resources in an attempt to obtain more accurate estimates. In addition, as pointed out by Renssen and Nieuwenbroek (1997), respondent burden is of concern, so multiple surveys may be used to ask more questions than can be done in one survey.

In this paper, two methods for combining information from surveys are discussed: multiple frame methods and multivariate small area models. Challenges in implementing these methods and future directions for research are also discussed.

2. Multiple Frame Surveys

In a multiple frame survey, several possibly overlapping frames cover the population of interest. González-Villalobos and Wallace (1996) describe multiple frame agricultural surveys used around the world; one frame is a list of agricultural holdings (cheaper to sample but incomplete), and the other is an area frame (complete but more expensive to sample). A multiple frame approach is also useful for locating and studying members of a rare population: a survey of diabetes patients in a registry may be supplemented by a general population health survey. In some cases it is possible to

identify duplications in the frames in advance, so that holdings in the list frame are excluded from the area frame; in that case, the separate estimates from the two frames are summed to estimate population totals. If the frames overlap, estimators discussed in Lohr and Rao (2000) may be used to combine information from the two frames; the pseudo-maximum likelihood (PML) estimator of Skinner and Rao (1996) allows using the same set of weights for all survey variables. The two-frame PML estimator of the population mean adjusts the sampling weights using the two survey estimates of the population size in the intersection of the frames.

We have extended the PML method to more than two frames by using estimates of the intersecting population sizes, and the variances of those estimates, for all subsets of the frames. The properties of the estimators are similar to those of the dual frame PML estimator.

The survey designs for the different frames are assumed to be independent. Multiple frame surveys also implicitly assume that the same questions are asked of the respondents sampled from each frame so that differences among estimates from the various frames are due to sampling error and the subset of the population in the frame, not question wording or mode of administration. Separate surveys with partially overlapping information can also be considered in the multiple frame setting if care is taken to distinguish questionnaire effects.

One advantage of multiple frame methods, which may lead to their increased use in the future, is that large national surveys may be easily supplemented with additional data from a state or province, or for a subgroup of the population. This contrasts with current practice in the U.S., in which most states wanting higher precision for estimates conduct their own survey with no coordination with federal efforts.

3. Small Area Methods

Surveys such as the U.S. Current Population Survey (CPS) give accurate estimates of poverty or unemployment at the national level. These surveys do not, however, contain sufficient sample sizes to give reliable estimates by themselves of "small areas" such as counties or minority groups. Current methods for estimating poverty in small areas incorporate auxiliary administrative information from sources such as tax records and food stamp programs as explanatory variables in a regression equation; the predicted value of the regression is combined with a direct estimate of poverty from the CPS to estimate the county poverty rate. This approach assumes that the administrative data are without errors and are known for all small areas; it does not incorporate information from other surveys or use longitudinal information.

Lohr and Prasad (2001) introduced an extension of the nested-error regression model of Battese et al. (1988), who used auxiliary satellite data to improve crop cover estimates from an agricultural survey. Here, we consider a situation in which the auxiliary data comes not from satellites or administrative records but from another survey. Suppose there are a total of t small areas, and that area i has N_i population units. Let y_{ij} denote a characteristic of interest for the jth observation unit in area i; y_{ij} is observed for a sample of units in one of the surveys. Let $\mathbf{x}_{ij}$ denote a vector of other characteristics for unit j of area i; $\mathbf{x}_{ij}$ is observed for a sample of units through a second survey. Let

$\mathbf{u}_{ij} = (\mathbf{x}_{ij}, y_{ij})$. For estimating income, y_{ij} is the income of household j in area i, measured in the CPS, and $\mathbf{x}_{ij}$ might be related quantities measured in another survey. In addition, there may exist various covariates from administrative records. Note that $\mathbf{x}_{ij}$ may actually be the same characteristic as y_{ij}, but measured in another survey under perhaps other conditions or at a different time; if $\mathbf{x}_{ij}$ and y_{ij} coincide, the dual frame setting applies.

Since different surveys are involved, the observation units sampled will often not coincide. Even when the surveys employ the same primary sampling units, the individuals sampled often differ. An approach is desired where overlap of the observation units can be used if it occurs, but where the second survey can provide auxiliary information even with no or little overlap.

We adopt a multivariate mixed unit-level model,

$$\mathbf{u} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{v} + \boldsymbol{\varepsilon},$$

where $\mathbf{u}$ is a vector of the $\mathbf{u}_{ij}$'s for units measured in either survey (some values may be missing because unit j is observed in only one survey), $\mathbf{A}$ and $\mathbf{Z}$ are known matrices, and $\mathbf{v}$ and $\boldsymbol{\varepsilon}$ are independent random vectors with mean $\mathbf{0}$ and respective covariance matrices $\mathbf{G}$ and $\mathbf{R}$. With this setup, the mean of the variable y in small area i is estimated using the best linear unbiased predictor under the multivariate model. The estimator depends on the means of $\mathbf{x}$ and y in the area if they are measured, borrows strength from other small areas through the model, makes use of administrative data through the regression parameters $\boldsymbol{\alpha}$, and uses the correlations between measures in the two surveys both at the small area and the individual level. If all observation units coincide for the two surveys, the multivariate estimator of Datta et al. (1999) follows as a special case.

There is no restriction on survey design, question wording, or mode of administration for this framework; precision of small area estimates is improved as long as the variables from the two surveys are correlated at the small area level. This model also provides information about multivariate relationships across surveys. To use it, however, it is necessary to be able to estimate the multivariate components of variance involved in $\mathbf{G}$ and $\mathbf{R}$, either from the survey data or other empirical work.

Several large U.S. surveys are now being redesigned so the linkage required for this method of small area estimation will be possible. The U.S. National Health and Nutrition Examination Survey is now linked with the National Health Interview Survey; the same primary sampling units, although not necessarily the same individuals, are in both surveys. In addition, both surveys are linked to administrative Medicare and National Death Index records.

4. Practical Problems and Challenges for the Future

All of the methods discussed above provide snapshots at a particular point in time. I believe the next theoretical challenge for this area of research is how to employ longitudinal data from multiple sources.

Multiple frame and small area estimation methods for combining survey information are valid theoretically and perform well in simulation studies. In practice, there are operational challenges to overcome. The PML estimator for multiple frames requires knowledge of frame membership for all

sampled units; if units sampled from one frame are mistakenly said to also belong to the other frame when they do not, the estimator may be biased. The sensitivity of the estimator to misspecified frame membership is unknown. Similarly, in the unit-level model for small area estimation described above, it is necessary to be able to match individuals across surveys if there is overlap. One option is for modified data collection procedures and processing that allow better and more accurate identification and matching of respondents.

With better identification and matching, however, comes increased concern about protecting the privacy of survey respondents. The ISI Declaration on Profession Ethics specifies that statistical inquiries should avoid "undue intrusion," and suggests making greater use of administrative data and linking records. However, more extensive linkage and combination of survey data may in fact lead to undue intrusion, because survey respondents may be unaware of the extent to which their data are used.

REFERENCES

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.

Datta, G.S., Day, B. and Basawa, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *J. Statist. Plan. Inf.*, **75**, 269-279.

González-Villalobos, A. and Wallace, M.A. (1996). *Multiple Frame Agriculture Surveys, Vols. 1 and 2.* Rome: Food and Agriculture Organization of the United Nations.

Lohr, S. and Prasad, N.G.N. (2001). Small area estimation with auxiliary survey data. Technical report.

Lohr, S. and Rao, J.N.K. (2000). Inference in dual frame surveys. *J. Amer. Statist. Assoc.*, **95**, 271-280.

Renssen, R.H. and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *J. Amer. Statist. Assoc.*, **92**, 368-374.

Skinner, C. J. and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. . *J. Amer. Statist. Assoc.*, **91**, 349-356.

RESUME

On examine des questions concernant la combinaison des données de plusieurs échantillons complexes. On considère l' estimation utilisant les informations tirées de bases multiples de sondage, et on propose un modèle multivarié de régression á erreur emboîtée afin d' obtenir des estimations pour des petites régions. Les problèmes pratiques sont illustrés avec des échantillons des Etats-Unis.