SMALL AREA ESTIMATION[1]

Ayah E. Johnson
2101 East Jefferson Street, Suite 500, Rockville MD 20852

1 Introduction

Researchers are often requested to come up with estimates for states, counties, and subpopulations of interest such as statutory-defined minority groups, in order to assist in the formulation of policy in areas such as welfare and health care. However, it is difficult to meet this need for information for each of the states or counties because of the prohibitive cost of conducting surveys which will yield reliable estimates at the national level, for each of the states, and within the states for the different subpopulations of policy interest. When using data bases created based on national surveys, standard methods of survey estimation for states or county breakdown because of one two reasons: (1) sample sizes, at the state or county level, are too small for reliability; or (2) sample sizes are large enough but the coverage of the population within the state is not adequate. For example, data from a national survey may not contain any sample data for a specific state, or the sample data may be primarily from an urban area with no or minimal data from rural areas.

Alternative methods of estimation for small areas (domains) have been and are being considered. A current Working paper prepared by the subcommittee on small area estimation (Statistical Working paper-Indirect Estimators in U.S. Federal Program, 1993) summarizes a special class of estimation methods referred to as small area or indirect estimators currently used by federal agencies. A comprehensive literature review of small area estimation techniques is provided by Purcell and Kish (1979), and in a monograph from the National Institute on Drug Abuse, Steinberg (1979). Such indirect estimators include the synthetic estimator (NCHS, 1968), the sample regression estimators, the poststratified estimator, and the composite estimators. (An extended list of references of studies investigating this issue is included). However, these estimators are considered to be either biased and/or have relatively small variances.

The common thread in all these approaches of small area estimation is the exploitation of symptomatic information collected from other domains, or from other surveys, and/or other time periods in conjunction with the use of one or more modelling techniques to try to estimate for the domain of interest. Symptomatic information generally includes aggregated measures which capture characteristics of the small domain. For example, estimating health care expenditures at the state level can be derived by estimating health care expenditures for the nation by various demographic characteristics, and applying this estimate at the local areas using census information on the demographic composition of that area. Another methodological approach, designed to increase the sample yield when repeated surveys are available, is combining data systems covering the different time periods. Their data can be accumulated, and estimates can be made using "all" available information. Malec, Sedransk and Tompkins, 1991 summarized some of the major concerns with "indirect estimators": (a) If the implicit assumption that the small domains resemble each other fails, the estimators may be biased; (b) the mean square error of indirect estimators is high, because of bias; and (c) the data accumulated for repeated surveys covers different time periods, thus point estimates are likely to be biased and difficult to interpret because they don't cover the same period.

One possible method which has not been explored is pooling several data systems which have been created based on different national surveys, rather than pooling symptomatic information. Pooling symptomatic information implies using logical demographic relationships in combination with statistical relationships based on other data to come up with required estimates (Purcell and Kish, 1979). For example, statistical equations are used to relate growth in the population to growth in symptomatic variables such as the number of births, deaths and migration in a given area. In contrast, pooling several data systems implies merging sampling units from two or more national surveys for the domain and the time period of interest, and incorporating them into one data system. In the example above, one would not use the symptomatic information from one state to infer behavior in a different state. Using the different national data systems, one would bring persons interviewed by two or more national surveys for the state and merge them into a single data system. The pooling of data systems will increase the effective sample size and the population coverage for each of the state. This is implicit in composite estimation where one uses sample estimates for small domains when available and combine them with synthetic estimates. In addition, pooling sampled units from two or more national surveys which are similar will, in effect, provide information on the sampling units (sampled persons) which are current, detailed and correlated for the statistic of interest. Therefore, it will capture the variation within the different domains and preserve the variation between domains. Once the sampling units from the different national surveys are merged into one data system, there will be a need to formulate the estimation strategy. Some of the same variables will be collected by the national surveys, and some will not. If the variable of interest, e.g. health care expenditures, is collected by one but not all surveys, then one needs to either impute or predict the value where it is missing.

Pooling of data systems based on national surveys does not preclude using other information which could enhance the estimation process. For example, to allow for estimation of health care use and expenditures, symptomatic information related to the overall domain of interest from other auxiliary data sources can be added. Thus, when making estimates of variables related to health care, one can use the Area Resource File System (ARF) which contains information on health facilities, health professions, health status, and socio-economic and environmental characteristics for each of the U.S. counties.

A set of surveys will be called "similar" if they have the same definition of a sampling unit, they cover the same time period, the survey design is comparable, and the survey instruments measure a reasonable number of the same variables. In this paper we outline a method that could exploit information from cross sectional national surveys with similar designs which were collected for the same time period with different objectives. One class of such surveys are the household surveys such as the Census Current Population Survey (CPS), the National Health Interview Survey (NHIS) and the National Medical Expenditure Survey (NMES). These three surveys use complex multistage sample designs, and all three surveys cover the civilian noninstitutionalized population. The elemental sampling unit for each of these surveys is a person within the household, and there is a common set of "years" for which data is available. Moreover, some of the basic demographic or socio-economic data is collected using similar survey instruments.

A general description of the methodology and the estimation process is described in section 2 -- we describe the process for pooling the data systems and deriving the estimate of interest for any "small domain". An application of a possible merge of the three data systems to derive estimates of health care use and expenditure at the state level using NMES, CPS and NHIS is described in section 3.

2 Method and Estimation

The art of population estimation is to make maximum use of all the data available by combining traditional and nontraditional sources of data. Suppose we wish to estimate a characteristic X, e.g. health care expenditures for physician visits, for each of the states within the U.S. A national survey which collected data on the variable X is available-- we call this survey the primary survey. Estimates for the characteristics X cross-classified by nonoverlapping and exhaustive subgroups of the population can be computed with the required reliability but only for larger domains defined at the level of the nation or at the level of the four major regions. Also, suppose that there are other surveys (secondary sources) with similar designs, for the same time period with the same sampling units but with different objectives. If one can effectively pool data systems from two (or more) national surveys, we will be able to increase the effective sample size, increase the coverage necessary to make estimates, and increase the reliability of the estimates for smaller domains. When combining the different data systems into one large sample, one has to standardize the information collected by each of the surveys and to account clearly for the sampling units in the joint sample.

Such an estimation process could be effectuated in the following manner:

1. ***Define the domains for which we need estimates of characteristic X.***

For example, estimates of health care expenditures for physician visits (X) are needed for each of the 50 states in the U.S. and the District of Columbia (a state is considered a small domain in this example.)

2. ***Identify those domains where the primary survey has sufficient coverage.***

When a domain is defined as a state, coverage can be defined as a minimum number of urban/rural PSUs and a large enough sample within these PSUs to meet the reliability requirements for the estimate of the characteristic X. The primary survey could be the NMES whose major objective is to provide unbiased estimates of medical use and expenditures. The second step is to identify those states where the primary survey has sufficient number of sampling units and has appropriate coverage to meet pre-specified reliability requirements. One can choose to do the supplementation without this step. It will basically increase the effective sample size even when there is sufficient information from the primary survey. If the primary survey does not have any sampling units, sampling units from the secondary surveys will be used.

3. ***Supplement the data system from the primary survey with additional sampling units and their respective data from the secondary survey(s). The auxiliary data will be at the sampling unit level; there is no record matching involved, thus we need not know the identity of the sampling unit.***

Since the surveys are conducted independently, the likelihood that a household will participate in more than one survey in a given year is small. Nevertheless, since each of the surveys used is a national survey, each person in the nation had more than one chance of selection. Therefore, we need to adjust for multiplicity.

Let n_s denote the sample size for survey s ($s=1,2,3,\ldots k$); and, let n_t denote the pooled sample size. The multiplicity adjustment can be defined as P_s, where:

$$P_s = n_s/n_t \qquad s=1,2,\ldots k.$$

if W_{sj} denotes the original sampling weight for the sampling unit from the s-survey, then the weight, W_{mj}, after the multiplicity adjustment will be:

$$W_{mj} = P_s * W_{sj}$$

In the example above, sampling units from both the CPS and the NHIS can be used to augment the NMES data file.

4. ***Derive and compute the sampling weight for each sampling unit within the new data system.***

After the data systems are pooled, the sampling weight, W_{mj}, associated with each sampling unit represents a surrogate measure of the sampling unit's probability of selection. Since the pooled data systems have similar designs and cover the same time period, one can recalibrate the sampling weight by using postratification. Thus, since W_{mj} is the weight associated with person j after all data systems have been incorporated, and the adjustment for multiplicity was done, one can define a weight W'_j so that

$$W'_j = A_j * W_{mj}$$

where A_j is the postratification adjustment. The sum of the poststratified weights W'_j can reflect the CPS totals for the appropriate year. This postratification process can, of course, be done by a set of prespecified weighting classes such as age, race/ethnicity, sex, poverty status, and geographic region. If the "small area" is a "state" one can use the number of persons within a state as the weight class.

The secondary data systems may have auxiliary data for each sampling unit, but not a reported measure for the variable of interest-- X. For example, socio-demographic and self-perceived health status and health care use is available for each sampled person in the NHIS sample. However, if X is a variable measuring health care expenditures, and the secondary surveys do not collect information on health care costs, the next step in the process is to:

5. ***Predict or impute the value for the characteristic X for each of the sampling units, using the relevant auxiliary information.***

Given the auxiliary information, one can use the sample regression method to predict the characteristic of interest for the variable X, or one can use an imputation technique, such as the "hot-deck" or multiple imputation, to impute the value of X for those sampling units (persons) where the information was not collected. Purcell, 1979, notes that the prediction using sample regression method seem to have the greatest potential and accuracy whenever good sample data on the variable of interest is available, and when there is a capability to build a good predictive model. On the other hand, model based and "hot-deck" imputation have been demonstrated to yield comparable estimates when imputing for missing data. Thus, the choice between model-based and hot deck imputation can be left for the analyst. (Multiple imputation techniques could be used to try to capture variance due to imputation.)

6. ***Estimate X for the small domains.***

Obtain estimates where for example X is yearly per capita health expenditures for each state for which, after pooling the data systems, has proper coverage and sufficient sample size to make reliable estimates. The estimate for the domain h can be expressed as:

$$\hat{X}_h = \frac{\overset{P}{\Sigma}\, W'_j * \hat{X}_{hj}}{\underset{Total}{\Sigma}\, W'_J} + \frac{\overset{S}{\Sigma}\, W'_j * \hat{X}_{hj}}{\underset{Total}{\Sigma}\, W'_J}$$

The first component of this sum is based on information collected for the primary survey (P).

The second component is based on data from the secondary data systems after having imputed or predicted X (S). Thus, the quality of the second component of the estimator is dependent on the quality of auxiliary sample data from the secondary sources, the timeliness and the relevance of data collected. It also depends on the quality of the predictive model (or imputation) used to estimate X for those sampling units where it is missing. If one assumes that bias due to imputation is zero or relatively small, then the estimator is unbiased since:

$$E(\hat{X}_h) = E\left[\frac{(\overset{P}{\Sigma} W'_j * \hat{X}_{hj}}{\underset{Total}{\Sigma} W'_J} + \frac{\overset{S}{\Sigma} W'_j * \hat{X}_{hj})}{\underset{Total}{\Sigma} W'_J}\right] =$$

$$= E\left[\frac{(\overset{Total}{\Sigma} W'_j * \hat{X}_{hj})}{\underset{Total}{\Sigma} W'_J}\right] = X_h$$

7. ***Obtain variance estimates.***

Since only data systems with similar designs are being pooled together, variance estimation which accounts for complex survey designs can be used. Moreover, since we corrected for multiplicity associated with the sample selection, the variance is additive. (One component of the variance which may be harder to compute is the variance due to imputation.) The Taylor Linearization approach is the simplest method to use for variance estimation since there is no requirement to create replicate weights.

8. ***Assess the quality of the estimates.***

Estimates and the respective variances based on accumulation of different data systems can be evaluated by doing various testing procedures. The first is to compare the estimates based on the pooled sample, to point estimates obtained for small domains where the sample yield, the coverage and the reliability requirements were met in the primary surveys. When state level estimates are available, either from states or from other sources, we can try to compare the estimates obtained using this new strategy with published (or otherwise available) estimates. Also, comparisons with other type of "indirect estimators" may also be performed.

In the remaining section we will illustrate the use of this procedure to estimate health care use and expenditure at the state level.

3 State Level Estimates of Health Care Use and Expenditure

Estimates of health care use and expenditures at the state level are not available because of the prohibitive costs of collecting sufficient data to yield reliable estimates. However, as noted above, three national household surveys collect person level information that includes measures and/or correlates of health care use and expenditures: The NMES, the CPS and the NHIS. Data systems created based on these three national surveys can be concatenated and an estimation strategy can be derived to come up with the much needed estimates. Once the data systems are concatenated, the total number of records will exceed a quarter of a million persons who participated in one of these surveys. Two methods are being considered for supplementing the measures of health care use, U, and the respective expenditure Y: an imputation and a model based approach. Characteristics of the point estimates of medical use and expenditure and their associated variance is then described.

Given the information that is available from the three data systems, the concatenation of the CPS, the NMES, and the NHIS can be done by merging the three persons level files with information that is common to the three data systems. Such information will include, for each sampled person, the socio-demographic and economic characteristics, the family relationships, health status, medicare or medicaid eligibility and health care use and expenditure when the national survey collects the information. In addition, information that is available on one but not all of the systems, but is relevant to health care use or expenditure, will also be included. This will be done with the anticipation that if the data is needed to attain an estimate of medical use and/or expenditure it will be imputed. By pooling all sampling units from the three data systems, we increase the effective sample size, and probably the coverage within any given state.

3.1 Health Care Use and Expenditure Data--Supplementing Using Imputation

Person level use and expenditure for medical services can be imputed using the hot-deck or it can be predicted using a modelling approach. The hot-deck imputation strategy is easy to implement, preserves the distribution, the percentiles, the measures of spread and the covariation in the data base. The basic idea behind the hot-deck imputation strategy is to identify predictors of health care use, cross classifying the data by these predictors to create a pool of donors and recipients with similar characteristics. A random "donor" is selected and its reported value is used to impute for the matched recipients. This method of supplementation of data is effective when the proportion of persons in the pool data system with reported health care use and/or expenditures is at least equal to the proportion of persons that do not have that information. The requirements for hot-deck imputation are usually more stringent with at least 20 donors per cell and at least a ratio of two donors per recipient.

3.2 Health Care Use and Expenditure-- Model Prediction

Predicting medical use and expenditures is complex. Newhouse and Phelps, 1976, and Duan, Manning, Morris, and Newhouse, 1981, described and compared alternative models for the demand and expenditure of medical care. The requirements by Newhouse and Phelps focus on: (1) disaggregation of medical services so that they are not considered a homogeneous commodity; (2) treatment of health insurance as endogenous; (3) permitting price to vary among providers and selecting the price of the provider selected as endogenous. Duan et al., 1981 requires a distinction among: (1) non-spenders for medical care; (2) spenders for ambulatory care; and (3) spenders with inpatient utilization.

One can adopt their recommendation and adopt their models to predict the demand for health care services given specific characteristics of sampled persons. It requires modelling separately the demand for each type of medical service: ambulatory visits, outpatient and emergency room visits and inpatient stays. The prediction of medical use and expenditure for each sampled person has to be done in three stages: (1) predicting the likelihood that a person will use medical services; (2) predicting the actual use of medical services, e.g. number of doctor visits in a given year; and (3) predicting expenditures for medical services used. As noted above, the models are dependent on the type of medical service sought: ambulatory care, emergency room visits, outpatient visits or inpatient stays. The objective for the first model is to predict the likelihood that a sample person will use medical services. A logistic model could be used to predict use or non-use of medical services by type of service. The second and the third models could be weighted least square models (WLS) as used by Duan et al., 1981. The dependent variable of interest is the use of medical services (number of visits to a physician, number of days in a hospital), and the set of the explanatory variables that are consistent with the demand equations are: age, race, education, health status, measures of disability, size of the city, employment status, region, health insurance coverage, medicare coverage, medicaid coverage, family income physicians per 100,000 in a county, and hospital beds per 100,000. As interaction terms one can include a measure that will capture the correlation among family members. The majority of these variables are reported by the CPS, with the exception of health status and the variables describing health resources within the area. The NHIS has extensive information on health status and disability and some information on use of medical services.

Additional information that characterize the local area in terms of its provision of health services could be obtained from the Area Resource files (ARF) and appended to the concatenated file. Expenditures for medical services are dependent on use, health insurance status, employment status, age, education, eligibility for medicaid, medicare. Again, given the sampled person characteristics and the type of services used, expenditures would be predicted for those sampled persons for which they were not collected.

The robustness of these models could be tested using the NMES data since both use and expenditures are collected for this survey. Moreover, these models could be tested for subsets of the NMES population. For example, one could find out whether there are regional differences which should be incorporated by evaluating the fit of these models for each of the nine census regions.

3.3 Point Estimates and Variance estimates

Given the methodology described above, once the imputation process (whether using the hot-deck or a modelling approach) is completed one can estimate the mean use and mean expenditure for medical services by type of service for states where the representation is adequate. As noted above, the estimators are unbiased.

Using the poststratified weights, one can estimate the variance of the estimators accounting for the complex nature of these three designs. (The variance of the estimator will incorporate the variance due to imputation if multiple imputation is used.) Since the person level information at the state level is available, both the between and the within state variation will be greater than the estimated variation when using a synthetic estimator.

4.0 Summary and Conclusions

This method of expanding the sample pool can be viewed as "borrowing sampling units" rather than "borrowing strength" from other data sets. Merging of data systems which were created based on national surveys is possible since these data are being disseminated through public use tapes. Moreover, rapid advancement in computer technology should allow for combining different data systems so as to improve local area estimation. Problems of data storage and computational speed have been alleviated. The cost of concatenating the data sets is not trivial, but it is certainly lower than conducting a full blown survey which will require sufficient sample size and coverage at the state (county) level to meet reliability requirements. One potential stumbling block is the confidentiality issues that might impede the use of this procedure. There is no need to get the identification of the sampling unit, but one might need state and a breakdown on metro/nonmetro areas within the state, or an identifier of the primary sampling units to supplement the primary data system to produce state level estimates.

In this paper we have focused on concatenation of data systems which were developed based on national surveys as a mechanism to increase the effective sample size and the representation of sampling units within states. When imputation is needed to supplement information that was not collected for the secondary surveys, but is needed for estimation, one can use not only the national data systems but information from additional sources. For example, for the medicare and the medicaid population it is possible to draw information on use and expenditures for medical services when predicting these variables from administrative records. Those are additional sources of data which could increase the reliability and the quality of the imputation and therefore improve the quality of the estimator X which is of interest. Thus one way to increase the quality of the small area estimators is to pool different data systems to one large national sample and to use additional sources of data to enhance the quality of the imputation process.

References

Duan Naihua, Manning Willard. Jr. G., Morris Carl N., and Newhouse, Joseph P., 1982. A Comparison of Alternative Models for the Demand for Medical Care. Prepared under a grant from the U.S. Department of Health and Human Services.

Elston, J., Koch, G., and Weissert, W. (March 1991) Regression-Adjusted Small Area Estimates of Functional Dependency in the Noninstitutionalized American population age 65 and over. American Journal of Public Health, 81 (3), 335-341.

Gonzalez, M., and Hoza, C. 1978. Small-Area Estimation with Application to Unemployment and Housing Estimates. Journal of the American Statistical Association, 73 (361) 7-15.

Heeringa, S., 1981. Small Area Estimation: Prospects for the Survey of Income and Program Participation: A Review of the Literature and Interim Report on the Methodology. Survey Development Research Center.

Heuser, R. et al. Synthetic estimation applications from the 1980 National Natality Survey (NNS) And the 1980 National Fetal Mortality Survey (NFMS), 31-51.

Landwehr, J., Pregibon, D., and Shoemaker, A. (March 1984) Graphical methods for assessing logistic regression models. Journal of the American Statistical Association, 79, (385), 61-73.

Makuc DM et al., 1991. Health service areas for the United States. National Center for Health Statistics. Vital Health Stat (2) 112.

Malec, D., Sedransk, J., and Tompkins, L., 1991. Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey. Presented at the workshop: "Bayesian Statistics in Science and Technology", Carnegie Mellon University.

Marker, D. (January 29, 1993) Small area estimation for the National Health Interview Survey. Presented to the Washington Statistical Society and the National Center for Health Statistics.

Newhouse, Joseph P. and Phelps, Charles E., 1980. New Estimates and Income Elasticities of Medical Care Services.

Pfeffermann, D., and Burck, L. (December 1990) Robust small area estimation combining time series and cross-sectional data. Survey Methodology, 16 (2), 217-237.

Purcell, N.R., and Kish, L. (June 1979) Estimation for Small Domains. Biometrics, 366-384.

Sarndal, C. (September 1984) Design-Consistent Versus Model-Dependent Estimation for Small Domains. Journal of the American Statistical Association, 79 (387), 624-631.

Schaible, W. et al., 1979. Small area estimation: An Empirical Comparison of Conventional and Synthetic Estimators for States. National Center for Health Statistics. Vital Health Stat (2) 82.

Wong, G., and Mason, W. (September 1985) The hierarchical logistic regression model for multilevel analysis. Journal of the American Statistical Association, 80 (391) 513-524.

Zeger, S., and Karim, M. (March 1991) Generalized linear models with random effects; a Gibbs sampling approach. Journal of the American Statistical Association, 86 (413) 79-86.

The Area Resource File (ARF) System. Information for Health Resources Planning and Research. Health Resources and Services Administration, Bureau of Health Professions, Office of Data Analysis, ODAM Report 7-89.

Statistical Working Paper: Indirect Estimators in the U.S. Federal Program, 1993. Office of Management and Budget.

Endnote

[1] The views expressed in this paper are those of the author and no official endorsement by the Department of Health and Human Services, or the Agency for Health Care Policy and Research is intended or should be inferred.